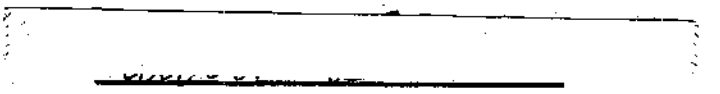
In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institution shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



CHARACTER RECOGNITION BY AREA MEASUREMENT

A THESIS

Presented to

The Faculty of the Graduate Division

by

Albert Whitman Bowers

In Partial Fulfillment

of the Requirements for the Degree

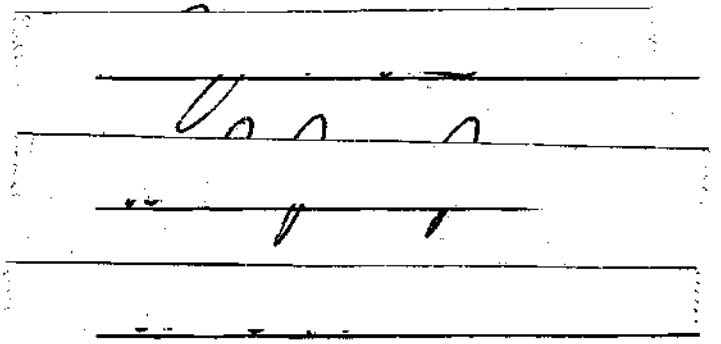
Master of Science in Electrical Engineering

Georgia Institute of Technology

June, 1964

3 D
12.12

CHARACTER RECOGNITION BY AREA MEASUREMENT



Date approved by Chairman: May 11 1964

ACKNOWLEDGMENTS

I wish to express my gratitude for the guidance given me by Dr. Benjamin J. Dasher as my thesis advisor. I thank my reading committee, Dr. William B. Jones, Jr. and Dr. William F. Atchison. I give special thanks to Dr. Irwin E. Perlin and Mr. Sterling P. Lenoir, Jr. for their assistance and advice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

SUMMARY

The field of character recognition may be described as the implementation of man's intuition and knowledge in an attempt to provide machines with the ability to decipher the writings of humans. Character recognition has evolved from man's attempts to simulate a specific function of the human brain. This function, known as perception, includes such human abilities as sight, hearing and touch. Because perception is difficult to understand and even harder to implement, its machine simulation has not progressed as rapidly as the simulation of other brain functions, such as addition, subtraction, and logical decisions.

The number of different approaches to character recognition is probably at least equal to the number of investigators in this field. However, each approach may be classified into one of two categories:

(1) template matching, or

(2) feature matching.

Template matching techniques classify characters by the comparison of a partial or a complete image of a single character with a similar image (partial or complete) of each character in the character set of interest. The images may be held for comparison on various media, such as a photographic transparency or a dot pattern stored in a digital computer memory. An array containing these images, or reference characters, is called a template. At least one image usually appears on the template for each character in the set. The physical size and required

resolution of these images restrict the number of characters which can be placed on a single template. For this reason, template matching techniques are usually restricted to character sets containing relatively few characters.

Feature matching techniques utilize a combination of characteristics such as cusps, closures, enclosed areas, etc., to classify each character. Their sequence of occurrence, or an otherwise unique combination of defined features, is used to classify each character. Generally, the number of features employed to classify a specific set of characters is less than the total number of characters in the set. Classification usually does not depend upon only one feature; therefore, the technique may be capable of handling higher degrees of distortion than a template matching technique using the same set of characters.

The object of the present study is to investigate the effectiveness of a character classification procedure for the classification of typed characters from the English alphabet, both upper and lower case, and the numerals zero through nine, excluding the numeral one which is identical to the lower case "L" in the type style investigated. Each character which is to be classified is extracted from a field of typed characters and depicted as a rectangular array of black and white dots. This dot pattern is converted by the procedure into a series of numbers representing the areas of the characters. The area measure is found by counting the black dots in the rows and columns of the array. The character is in black on a white background. The series of numbers for each character is called the character vector. The proper classification

of the character is achieved by comparing the character vector with each of the character vectors which have been stored to represent the characters in the set. The appropriate classification is selected on a best match basis.

The effects of distortion are considered in this analysis by applying a distortion factor to each of the elements in the character vector separately. The distortion factor, a constant, is used to define an upper and a lower limit for each element in the stored character vectors. The vector pairs, formed in this manner, define a solid in the vector space. Any region in this vector space which is common to two or more characters represents an ambiguity and signifies that the characters in question cannot be distinguished when they produce vector representations in this region. This method of treating distortion is used primarily as a measure of the similarity between characters when subjected to this classification procedure.

The item to be classified is positioned in the lower left corner of a 24 x 24 rectangular matrix. With this method of positioning it was found that only 15 rows and 20 columns actually contribute significant classification information. The encoding process was found to emphasize long straight character segments or segments with only slight curvature. Features with excessive curvature were seldom found to possess significant classification information. Thus the encoding process converts the characters into a rather crude block form and utilizes only those features which are converted into relatively long straight segments.

CHAPTER I

INTRODUCTION

The field of character recognition may be described as the implementation of man's intuition and knowledge in an attempt to provide machines with the ability to decipher the writings of humans. Character recognition has evolved from man's attempts to simulate a specific function of the human brain. This function, known as perception, includes such human abilities as sight, hearing and touch. Because perception is difficult to understand and even harder to implement, its machine simulation has not progressed as rapidly as the simulation of other brain functions, such as addition, subtraction, logical decisions, etc.

Character recognition is actually two distinct processes, character identification and character classification. Character identification is the process whereby a single character is found and extracted from a field of characters. The process of assigning the appropriate name to the character is called character classification.

Initial attempts to develop character recognition devices circumvented some of the problems of character classification by placing constraints on the character font. These constraints were imposed to make the characters compatible with a particular classification procedure. The type font used by Farrington Electronics, Inc. termed "SELFCHEK" is a typical example (1). This font constrains the character

to straight segments only and uses some built-in features (such as special serifs) which force otherwise similar characters to possess distinct differences. Another example is the magnetic ink reading process used on bank checks. The characters employed in this process are not easily readable by a person not acquainted with them since they have little resemblance to the English alphabet. The character classification in this system is accomplished by synthesizing the electrical signals from several parallel magnetic reading heads (2).

Other early machines have been built to read a limited number of characters that are not stylized. These machines were usually restricted to the numerals zero through nine and a few special characters. The IBM 1418 character reader is a typical example (3). The 1418 can classify twelve characters, the numerals zero through nine and two special characters. The twelve characters are in the type font used by most of the printing equipment employed by IBM's on-line printers for their digital computing machinery.

Recently character readers have been developed which can read and classify 64 characters (and in some cases more than 64 characters). Recognition Equipment, Inc. has advertised a machine which can read characters printed from standard typewriters (4). In general, character readers have been restricted to reading characters from only one type font on any single pass across a document. A subsequent pass over the document can be used to read a second font. This process can be repeated as often as necessary to read all of the fonts on the page. This type of multi-font reading capability can be achieved by interchanging character templates (a storage media containing an image of

each character in the set) or by making some modification to the decision logic (the classification process). These processes will be described in more detail later.

The near future is not likely to see a perfectly general character reading device; therefore, a finite number of character classifications will be a restriction for any realizable reading machine. However, the number of characters in the set can certainly be increased far beyond the current limit (approximately 64) before the finite limit is reached. Before the size of the character set can be appreciably extended techniques must be developed which compress more information about the character features into fewer bits of digital information. Current digital equipment techniques, employing magnetic drum and magnetic core memories, cannot handle much larger sets using coding techniques now employed. Certainly the future will bring better memory devices. Therefore, some compromise will probably result between better memories and better coding techniques. The ultimate character reader will probably follow the lines of a Perceptron (5) and utilize advanced memory devices.

At present, the number of different approaches to character recognition is probably at least equal to the number of investigators attempting to solve the problem. This fact is attributed to the lack of a clear-cut set of rules governing character identification and classification. Even though the number of techniques for character recognition is large, each can be classified into either one of two categories as proposed by O. G. Selfridge (6):

(1) a template matching technique, or a

(2) feature matching technique.

Template matching techniques usually display the character to be classified in its entirety on some media such as a photographic transparency or a dot matrix, as formed in a core memory. In this form the image is compared, either in its entirety or at specific points, against a similar representation for each character in the set. When the comparison is made at specific points the character set need not be stored explicitly but may be implied within the decision logic. Techniques of this type are valuable because of their simplicity. Changes in the character set are accommodated by interchanging the character template, or by modifying the decision logic.

Template matching techniques have several disadvantages. They are usually susceptible to variations in registration and to character distortion. Changes which are evident in a character from one appearance to the next could cause trouble for a template matching technique were it not for the ability to make classifications on a best match basis rather than an exact match. The results of the classification decision must indicate a single character designation, with some degree of assurance; otherwise, no classification is made and the document is rejected. Some machines have the ability to present the distorted character to the machine operator for classification, thus eliminating some of the rejected documents.

An example of a technique which employs template matching is found in a device proposed by W. J. Hannan of RCA (7). The device consists of three major units: a flying-spot scanner, an optical tunnel, and a memory unit. The positions of the lines and characters are

located by means of an electronic search of the document with the flying-spot scanner. Classification is by optical correlation. An optical "tunnel" is used to make a multiplicity of correlations simultaneously. The basic memory of the machine is a photographic plate, known as the correlation mask, which may be changed to classify any font.

The flying-spot scanner initially picks up the image of a character from a printed document. The image is transferred to the face of a dual beam oscilloscope. The dual beam device produces both a positive and a negative image of the character on the scope face. These two images are projected down the optical tunnel. The optical tunnel is a rectangular tube with each of the four inside surfaces mirrowed. These mirrowed surfaces cause the images from the oscilloscope face to be reproduced many times when they emerge from the opposite end of the tunnel. The photographic template is located opposite the light tunnel. This template contains a positive and a negative image of each character in the set. A pair of images from the light tunnel coincides with each pair of images on the template. A photocell, located behind each image on the template, receives the light from the images produced by the oscilloscope and light tunnel after it passes through the template.

The oscilloscope images are composed of a series of raster scans formed as the flying-spot scanner reads the character from the document. As the character is scanned, electronic integrators produce a cross-correlation between the document character and the image pair on the template which correspond to this character. In order to minimize

incorrect classifications a threshold value is established for the cross-correlation. If none of the cross-correlations exceed the threshold no classification is assigned to the document character. If more than one cross-correlation exceeds the threshold, again no classification is assigned to the document character. Experimental results indicate that five errors per million characters read can be expected from the device.

This device, although versatile (can classify a 91 character alphabet), has one major disadvantage as a multiple character reading machine. The number of character classifications which can be made by this machine during a single reading pass over a document is limited by the physical size of the character template. For each new character classification, the addition of an electronic integrator circuit is also required. Thus the cost of the machine is proportional to the number of characters in the set.

Feature matching techniques attempt to isolate a few distinguishing character features, such as cusps, closures, enclosed areas, moments, etc., which can be used to classify the characters in the set. These features are selected to suit the particular technique to be used for classification. Usually there are many features that can be used by any particular feature matching technique. The classification ability of each feature is usually evaluated by some appropriate procedure and the features which are most efficient are selected for use. In general, more than one feature will be required for the classification of a single character. The advantage of feature matching techniques is that the number of features required to classify all of

the characters is usually less than the total number of these characters. The cost of a feature matching character reader is, therefore, not necessarily proportional to the number of characters to be classified. In addition, since classification need not hinge strongly on any one feature, the technique may be capable of handling higher degrees of distortion than a template matching technique using the same set of characters.

A feature matching technique can be distinguished from a template matching one by its use of the relative position of the character features rather than their absolute position. When relative size can be incorporated with the relative position of the features, the feature matching technique becomes a versatile tool as a multi-font reading device. Relative size is usually harder to incorporate into a template matching technique than it is for a feature matching technique. This is evidenced by the fact that cusps, closures, parallel lines, etc., are effected very little by size.

An example of a feature matching technique is found in the classification procedure developed by Franz L. Alt of the National Bureau of Standards (8). This technique uses the first several moments of a character as its classification. The moments as used in this procedure are independent of the character location and any variations due to squeezing or stretching. The moments are made independent of character location by normalizing them with respect to the center of gravity of the character. The moments themselves are inherently independent of squeezing and stretching effects.

The moments of the characters are found from the approximation:

$$M_{jk} = c\Sigma\, x^j\, y^k$$

where the summation is taken over all "black" cells, c is the area of one cell in the place and x,y are the coordinates of some point in the cell.

The zeroth moment is the black area of the character and is represented by M_{00}. The center of gravity has coordinates:

$$\overline{X} = M_{10}/M_{00} \text{ and } \overline{Y} = M_{01}/M_{00} .$$

The remainder of the moments in the analysis are referred to these coordinates as a base to make the moments invariant under displacement.

The total number of moments required to classify all of the characters in a particular set is dependent upon the set. The largest set of characters studied consisted of the 26 capital letters and the numerals one through nine. The first six partial moments were required for this classification. The effects of distortion have not been fully studied for this procedure, but it is felt that a reasonable amount of distortion can be accepted.

The technique is well suited to the classification of characters from different type fonts having similar character shapes, because the procedure is insensitive to size and local feature changes within a character. The procedure is, however, sensitive to the global distribution of the black area over the character. Thus, it will not perform successfully on characters which are distinguished only by a round corner rather than a sharp corner, etc. When distortion appears as random addition and subtraction of small amounts of area over the

entire character the procedure can successfully accomplish classification. Heavy smears or large areas which are filled in or omitted would render this procedure ineffective. The principle advantage of the technique is that a small number of moments are required to classify a much larger number of characters.

The majority of the character recognition techniques that have become actual hardware devices are template matching techniques. The features that are used are usually quite general and hard to exactly define. For this reason the implementation of the classification procedure is necessarily elaborate. Usually, the more specific the features, the less effective is the classification procedure, as with the one described above, but the simpler will be the implementation of the procedure. These are probably the reasons why there are more template matching techniques than feature matching techniques.

The character classification procedure described in this thesis is a template matching technique which utilizes some of the advantages of feature matching. The classification process uses the relative position of the black area within a row or column of a rectangular matrix, on which the character has been superimposed. The black area within the rows and columns of the matrix are used to classify the character. The total number of rows and columns which are required for the classification is approximately one half of the total number of characters in the set. Thus, the procedure resembles a feature matching technique in that some degree of relative position is used and the number of items required to classify the characters is less than the total number of characters in the set.

CHAPTER II

CHARACTER RECOGNITION PROCEDURE

Objectives

The objective of this study is to investigate the effectiveness of a pattern classification procedure. The procedure, as described in the following sections of this chapter, is used to classify characters from the English alphabet, both upper and lower case, as well as the numerals zero through nine, excluding the numeral one which is identical to the lower case "L" in the type font investigated. A set of typed characters was obtained for study from an IBM Executive typewriter having a standard elite type font (IBM font number 02) and a carbon ribbon. The typed characters were clean and relatively free from character defects such as blurs or missing features.

The effectiveness of the classification procedure was evaluated using several degrees of character distortion. Distortion, in general, is a function of many parameters such as character shape, ribbon life, paper texture, and typewriter alignment. In order to consider distortion without specifying a statistical model or generating a large number of samples of each character, certain simplifying assumptions were made concerning the effects of distortion on the classification procedure. These assumptions are discussed in another section of this chapter.

A second objective of this study is to present a method for

finding a single character in a character field, and then encode it for presentation to the classification process. The encoding is performed in a manner requiring less digital storage than is required to store a dot-pattern replica of the character. This form of presentation as an input to the classification process is desirable in order to reduce the cost of implementation.

The Encoding Process

The evaluation of the classification procedure was performed on the Burroughs 220 digital computer. Before the evaluation could be performed the typed characters had to be transcribed into a form which is acceptable to the digital computer. Lacking means for automatic scanning of the typed characters, the encoding process was performed by hand. The character encoding consisted, in brief, of positioning each character accurately within a rectangular matrix and then counting the number of black squares.

Accurate registration of the characters during encoding was aided by typing a slash mark before and after each character. Figure 1 shows the entire character set as it appeared for encoding. The slash mark was used for alignment because it had the greatest vertical excursion, both above and below the line of type, of any character in the set.

/a/b/c/d/e/f/g/h/i/j/k/l/m/n/o/p/q/r/s/t/u/v/w/x/y/z/
/A/B/C/D/E/F/G/H/I/J/K/L/M/N/O/P/Q/R/S/T/U/V/W/X/Y/Z/
/0/1/2/3/4/5/6/7/8/9/

Figure 1. The Character Set as Presented for Hand Transcription

The registration and encoding of a character is accomplished in the following manner:

(a) A photographic enlargement is made of each character and its associated slash marks. The enlargement is made such that the total vertical height of the slash marks measures approximately 2.4 inches.

(b) A sheet of graph paper (10x10 grids to the half inch) having sufficient transparency for tracing purposes is positioned over the character as follows:

(1) A rectangle 2.4 inches high by 1.2 inches wide is drawn on the paper such that it coincides with the graph grids.

(2) The left edge of this rectangle is placed tangent to the left extremity of the character.

(3) The bottom edge of the rectangle is placed tangent to the lowest extremity of the character.

(4) Rotational alignment is achieved by positioning a line connecting the tops of the slash marks on each side of the character parallel to the top of the rectangle while maintaining the previous tangential requirements.

Figure 2 shows the letter "a" positioned within a rectangular matrix as described above. The width of the rectangle is slightly wider than the width of the widest character in the set and the height of the rectangle is equal to the height of a slash mark. The left edge and bottom edge of the characters were selected arbitrarily as the basis for accurate alignment. Another selection might produce different and perhaps better results from the classification procedure

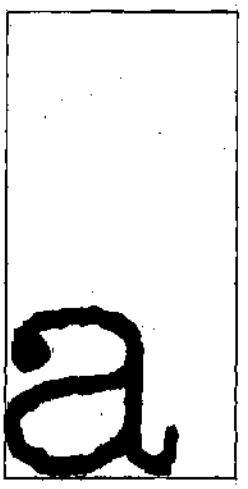
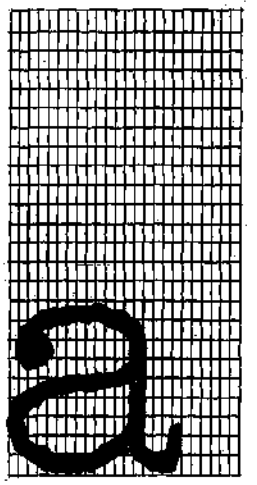


Figure 2. A Properly Registered Letter

since the basis for comparison of the characters in the classification procedure (to be discussed later) is predicated upon the position of the character within the rectangular matrix.

An arbitrary selection of 24 vertical divisions and 24 horizontal divisions was used as the encoding resolution. With this resolution the width of a vertical line segment on a character is equivalent to approximately two horizontal divisions and the width of a horizontal line segment is equivalent to approximately one vertical division. When a letter is blocked out in a rectangular matrix with this degree of resolution all of its characteristic features and most of its font pecularities are discernible.

After a character has been positioned in a rectangular matrix, as described above, it is encoded into a form which is acceptable to the digital computer. The particular coding method employed for this study is to represent each character by a 48-element vector in which each element of the vector represents the amount of black area contained in a row or column of the rectangular matrix corresponding to that element. The 48 elements are formed from the 24 rows and the 24 columns of the matrix. The amount of black area in a row or column is found by counting the number of black squares in the row or column. Figure 3 shows the letter "a" enclosed in a rectangular matrix and the 48 element vector representing the encoded character. The vectors for each of the 61 characters in the set used for this analysis are shown in Table 1.



	Horizontal Scanning	Vertical Scanning
a	11 16 07 05 11 08 06 07 11 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	02 06 06 06 06 04 04 05 05 05 03 03 05 09 09 02 02 02 01 00 00 00 00 00

Figure 3. The Letter "a" Correctly Registered and its Ideal Vector

Table 1. Ideal Character Vectors

	Horizontal Scanning	Vertical Scanning
a	11 16 07 05 11 08 06 07 11 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	02 06 06 06 06 04 04 05 05 05 03 03 05 09 09 02 02 02 01 00 00 00 00 00
b	10 05 04 04 04 04 05 12 06 02 02 07 00 00 00 00 00 00 00 00 00 00 00 00	01 01 01 01 01 12 12 03 02 02 03 03 03 02 03 03 05 04 00 00 00 00 00 00
c	03 11 07 02 03 05 06 08 13 06 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 05 07 03 02 03 03 04 04 04 03 05 06 06 05 03 06 00 00 00 00 00 00 00
d	13 06 04 04 04 08 07 12 07 02 02 09 00 00 00 00 00 00 00 00 00 00 00 00	03 05 05 03 02 03 04 04 04 04 04 06 12 09 01 01 01 01 00 00 00 00 00 00
e	08 05 02 02 17 04 04 12 05 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 05 04 03 02 03 04 04 04 04 03 02 04 04 03 01 00 00 00 00 00 00 00 00
f	14 02 02 02 02 02 02 13 16 06 06 10 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 03 11 11 05 04 04 04 04 04 02 02 04 04 03 00 00 00 00 00 00
g	07 15 04 04 08 13 06 10 04 04 05 16 08 00 00 00 00 00 00 00 00 00 00 00	04 05 07 07 04 05 06 07 07 07 06 05 07 06 05 04 06 06 01 00 00 00 00 00
h	17 04 04 04 04 04 07 11 08 02 02 06 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 12 12 02 03 03 01 02 02 03 03 02 06 07 01 01 01 00 00 00 00
i	16 02 02 02 02 02 02 08 08 04 05 05 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 03 05 06 12 10 04 01 01 01 01 01 01 00 00 00 00 00 00 00 00
j	09 04 06 06 04 02 02 02 02 02 08 00 03 05 03 00 00 00 00 00 00 00 00 00	02 04 05 03 01 03 03 03 03 06 07 14 11 00 00 00 00 00 00 00 00 00 00 00
k	18 05 04 05 06 05 06 10 10 02 02 05 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 12 11 03 02 02 01 05 06 05 05 05 04 04 03 01 01 00 00 00 00
l	17 02 02 02 02 02 02 02 02 02 02 09 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 02 02 02 12 12 07 01 01 01 01 01 01 01 00 00 00 00 00 00 00
m	18 07 07 07 07 06 08 17 11 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 09 09 03 02 01 01 02 08 08 08 02 01 02 02 03 08 08 01 01 00 00
n	18 04 04 04 04 04 06 14 08 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 09 07 02 03 02 01 02 03 03 02 02 08 07 01 01 01 00 00 00 00
o	07 06 05 04 04 04 04 05 09 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 05 04 02 03 02 02 02 02 02 02 02 02 03 05 05 00 00 00 00 00 00 00 00
p	12 02 02 11 02 06 04 04 05 05 11 12 00 00 00 00 00 00 00 00 00 00 00 00	02 04 04 04 02 12 12 07 04 04 02 02 01 01 01 03 04 05 05 03 00 00 00 00
q	11 04 01 01 09 05 04 03 03 04 05 11 06 00 00 00 00 00 00 00 00 00 00 00	04 06 02 02 02 03 03 04 04 03 03 04 05 13 04 02 02 01 00 00 00 00 00 00
r	12 02 02 02 02 05 08 16 08 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 03 09 03 03 02 03 02 02 01 01 03 04 03 02 02 00 00 00 00 00
s	12 07 04 06 10 04 04 08 09 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 07 05 03 05 03 03 03 03 03 03 04 03 07 03 00 00 00 00 00 00 00 00 00
t	06 10 04 04 02 02 02 02 16 02 02 02 00 00 00 00 00 00 00 00 00 00 00 00	01 01 01 01 01 01 01 11 11 03 03 03 03 03 03 02 03 02 00 00 00 00 00 00
u	13 07 05 04 04 04 04 11 11 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 08 09 01 01 01 01 01 01 03 03 04 09 09 01 01 01 01 00 00 00
v	03 04 04 04 05 04 04 17 17 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 02 03 05 05 05 03 02 03 03 04 05 04 02 02 02 02 02 00 00 00
w	02 04 08 08 07 05 04 04 16 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	01 01 01 01 03 05 05 05 03 02 02 02 02 02 03 04 04 04 04 02 01 01 00 00
x	16 04 06 06 02 06 06 15 14 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 04 06 05 05 02 03 03 02 06 05 05 04 03 03 03 03 08 00 00 00
y	05 09 06 05 01 02 02 04 04 04 17 17 00 00 00 00 00 00 00 00 00 00 00 00	02 05 05 06 05 05 06 06 06 05 03 01 02 04 05 04 02 02 02 02 02 00 00 00
z	15 06 06 03 03 06 06 14 14 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00	05 06 06 05 05 05 05 04 05 05 05 04 05 05 03 00 00 00 00 00 00 00 00 00

A	16 04 04 12 11 04 04 04 04 05 14 04 00 00 00 00 00 00 00 00 00 00 00 00	01 01 01 03 04 04 05 05 05 04 02 04 04 05 05 04 04 03 01 01 01 01 00 00
B	18 05 05 05 06 12 12 05 05 05 16 16 00 00 00 00 00 00 00 00 00 00 00 00	02 03 03 03 03 12 12 05 05 05 05 05 05 05 06 09 04 06 03 00 00 00 00 00
C	09 06 06 05 02 02 02 05 06 06 09 11 02 00 00 00 00 00 00 00 00 00 00 00	07 08 01 03 04 02 02 02 02 02 02 03 03 07 08 07 00 00 00 00 00 00 00 00
D	16 05 05 04 05 04 04 04 04 04 15 15 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 03 12 12 03 03 03 03 03 03 03 03 02 03 07 07 03 00 00 00 00
E	20 04 04 06 04 09 09 06 06 04 20 20 00 00 00 00 00 00 00 00 00 00 00 00	03 05 03 03 03 12 12 05 05 05 05 05 09 09 03 03 03 03 09 09 00 00 00 00
F	11 02 02 04 04 09 09 07 07 05 20 20 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 03 12 12 05 05 05 06 04 04 08 02 02 02 02 05 05 05 00 00 00
G	12 06 05 05 11 11 02 04 04 05 08 12 00 00 00 00 00 00 00 00 00 00 00 00	05 07 05 03 03 03 02 02 02 02 02 05 06 07 10 11 04 02 02 02 00 00 00 00
H	20 04 04 04 04 13 13 04 04 04 19 20 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 12 12 05 05 05 05 02 04 05 05 05 12 12 03 05 03 03 00 00 00
I	17 17 03 03 03 03 03 03 03 03 17 17 00 00 00 00 00 00 00 00 00 00 00 00	04 04 04 04 04 04 04 13 13 13 04 04 04 04 04 04 04 00 00 00 00 00 00 00
J	11 08 08 08 07 03 03 03 03 03 14 14 00 00 00 00 00 00 00 00 00 00 00 00	03 04 05 05 04 01 03 03 03 03 04 04 12 11 10 02 02 02 02 02 00 00 00 00
K	18 05 06 06 05 08 08 06 06 06 08 18 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 12 12 12 05 05 02 02 04 03 07 05 06 05 04 02 02 02 02 02 00 00
L	20 05 05 05 05 03 03 03 03 03 03 13 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 02 12 12 12 03 03 03 03 03 01 01 01 01 01 05 05 00 00 00 00
M	16 04 07 07 09 09 16 10 09 09 12 12 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 12 12 04 04 04 03 03 03 03 04 03 03 05 04 09 12 12 02 02 01 00
N	13 06 06 06 06 07 07 07 05 07 08 18 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 12 12 03 03 05 03 04 03 03 03 04 09 12 12 01 01 01 00 00 00
O	07 11 05 04 05 05 05 05 04 04 07 10 07 00 00 00 00 00 00 00 00 00 00 00	04 06 09 04 04 03 03 04 04 04 04 04 04 03 02 03 03 05 05 00 00 00 00 00
P	12 03 03 03 03 13 07 05 05 06 18 15 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 12 12 12 04 04 04 04 04 03 03 03 03 04 09 04 00 00 00 00 00
Q	05 04 07 11 06 10 11 04 04 04 04 05 06 08 08 00 00 00 00 00 00 00 00 00	03 07 05 03 05 06 05 04 04 04 04 05 05 05 07 06 08 04 04 03 00 00 00 00
R	15 04 04 04 04 05 13 04 04 04 09 17 00 00 00 00 00 00 00 00 00 00 00 00	02 02 03 03 12 12 05 03 03 03 02 02 03 03 04 05 09 07 02 01 01 00 00 00
S	07 15 06 05 05 05 15 12 05 05 06 13 10 00 00 00 00 00 00 00 00 00 00 00	08 11 10 06 05 05 06 06 06 06 05 05 04 09 07 04 00 00 00 00 00 00 00 00
T	14 14 02 02 02 02 02 08 08 08 08 20 20 00 00 00 00 00 00 00 00 00 00 00	06 06 06 04 04 04 04 04 04 13 13 04 04 04 04 04 04 06 06 06 00 00 00 00
U	06 11 06 05 06 06 06 06 06 06 06 18 18 00 00 00 00 00 00 00 00 00 00 00	02 02 02 09 11 12 04 03 04 02 02 02 02 04 03 04 12 11 10 02 02 02 00 00
V	04 04 04 05 04 04 05 06 04 06 10 18 00 00 00 00 00 00 00 00 00 00 00 00	01 02 02 02 03 05 05 06 05 04 02 03 04 05 05 05 05 03 02 02 02 01 00 00
W	07 07 09 09 08 09 09 08 04 04 17 17 00 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 05 09 10 06 04 05 05 03 03 04 03 06 05 09 09 07 02 02 02 01
X	17 04 05 05 05 04 03 05 06 05 12 18 00 00 00 00 00 00 00 00 00 00 00 00	03 03 03 03 04 06 06 06 05 04 02 04 06 06 06 04 03 03 03 03 00 00 00 00
Y	12 12 02 02 02 02 02 04 06 06 04 15 15 00 00 00 00 00 00 00 00 00 00 00	02 02 02 02 02 05 06 06 06 04 06 04 04 04 07 06 04 02 02 02 00 00 00 00
Z	20 06 06 06 07 03 03 06 06 05 16 16 00 00 00 00 00 00 00 00 00 00 00 00	01 01 07 08 08 05 05 05 05 06 06 05 05 04 05 04 03 07 05 05 00 00 00 00

0	08 04 05 04 04 04 04 04 04 04 04 08 00 00 00 00 00 00 00 00 00 00 00 00	04 08 04 04 02 02 02 02 02 02 02 02 02 02 02 04 07 04 00 00 00 00 00 00
2	17 02 03 03 05 07 02 04 05 04 05 09 00 00 00 00 00 00 00 00 00 00 00 00	05 06 06 03 05 08 03 04 04 03 03 03 03 03 05 04 00 00 00 00 00 00 00 00
3	02 08 03 02 02 02 02 02 02 04 10 03 02 02 15 00 00 00 00 00 00 00 00 00	03 03 02 03 04 03 04 04 04 04 04 04 04 03 04 05 03 00 00 00 00 00 00 00
4	02 03 02 19 17 04 04 04 04 04 02 02 02 02 02 01 00 00 00 00 00 00 00 00	01 02 03 06 11 09 02 02 02 02 02 03 10 10 02 02 02 02 01 00 00 00 00 00
5	04 04 04 02 02 02 02 02 02 05 12 06 02 02 03 15 00 00 00 00 00 00 00 00	08 08 04 04 04 04 05 04 03 04 04 08 05 01 00 00 00 00 00 00 00 00 00 00
6	04 10 04 04 04 04 05 08 11 04 03 02 03 03 06 04 00 00 00 00 00 00 00 00	07 07 04 03 04 04 03 05 04 04 04 03 03 04 04 09 06 01 00 00 00 00 00 00
7	02 02 02 02 02 02 02 02 02 02 02 02 05 05 18 00 00 00 00 00 00 00 00 00	08 08 03 01 01 01 01 01 08 10 04 08 03 03 02 02 02 01 00 00 00 00 00 00
8	08 04 04 04 04 04 04 12 11 04 04 04 05 11 05 00 00 00 00 00 00 00 00 00	03 05 07 06 04 04 04 05 05 05 05 05 04 03 06 06 07 04 00 00 00 00 00 00
9	05 03 03 08 02 03 02 13 06 05 04 04 04 04 09 00 00 00 00 00 00 00 00 00	05 06 08 03 02 03 03 03 03 03 04 08 04 05 03 06 07 04 00 00 00 00 00 00

Table 1. Ideal Character Vectors

The Scanning Process

The vector representation of a character is of little value unless it can be obtained automatically should the procedure be implemented. For this reason the following section is devoted to outlining a method which could be implemented to perform the encoding process. The procedure outlined here assumes that a television type raster scanning device similar to a TV camera is used to scan the characters.

The process of finding a single character and subsequently encoding it must be preceded by a procedure which selects an individual character from a field of characters. The following three operations are given as a means for isolating a single character in a character field.

(a) A horizontal raster scan of the document is made beginning at the top of the document. This scanning operation is terminated when a scan is made which contains an indication that black area has been detected.

(b) The raster scan is changed to a vertical scanning mode and the length of the raster line is reduced to equal the height of a line of print. Step (a) above has defined the top of a print line having a constant width which is preset into the machine. The vertical scanning is performed by scanning successive rasters proceeding from left to right. The vertical raster scanning is terminated when a scan is made which contains black area.

(c) The raster scan is changed at this point to horizontal scanning. At the same time the length of each raster is halved. The raster scanning begins at the bottom of the line and proceeds upward

until a scan is made which encounters some black area.

At this point the character can be considered as being correctly positioned within a hypothetical rectangle, Step (b) having defined the left extremity of the character and Step (c) the lower extremity. In order for the scanning process described above to be effective it must be assumed that each line of print on the document is separated from its neighbors, both above and below, by some small region in which there is no portion of any character from either of the lines, i.e., there is no overlap of the bottom of a character from the line above with the top of a character from the line below it. The line of print can be defined under these conditions as existing between the limits defined by the scan which detects the top of a character in the line (Step (a) above), to be called scan number K and scan number K plus 24. This procedure also assumes that a line of print will never be encountered which does not have at least one character which extends upward to the top of the line of print, i.e., either a capital letter, a lower case letter such as "b", or a number.

The character is encoded by using vertical raster scanning of the line of print, i.e., scanning from the bottom of the line to the top on each raster scan with each successive scan proceeding from left to right. The right edge of the character is determined when two or more successive scans do not contribute any additional black area. Normal typing sometimes does not exhibit sufficient spacing between characters for more than one scan to be obtained without detecting some of the black area from the adjacent character. If this problem is encountered then it will be necessary to define the right extremity of

the character in absolute terms such as a fixed number of scans. The classification procedure will then be modified slightly to allow for the additional area contributed by the next character in the line.

Each of the vertical scans correspond to the columns of the hypothetical matrix. The vertical scans are also divided electronically into segments each of which correspond to a row of the matrix. As the first column is scanned, the scan segments which are found to contain a portion of the character (black area) are counted and recorded as the value for vector element number 25. For each segment which contained a portion of the character, the count in the vector element counter corresponding to that matrix row is increased by one. In this way it is not necessary to store the image of the character in a digital memory before encoding.

To illustrate this process consider the encoding of the lower case "a" in Figure 3. The character is registered, as described above, in the lower left corner of the matrix. As column one is scanned two segments are found to contain black area. These segments correspond to rows three and four. At the completion of the scan for column one the count in vector element 25 is two, corresponding to the area in column one, and the counts for elements three and four are each one, corresponding to the area found in segments three and four of the first scan.

The second column is scanned and six segments are found to contain black area. The results of this scan produce a count of six in vector element 26 and a count of one for elements two, five, seven, and eight with elements three and four receiving an additional count

each to bring their individual totals to two. This process is repeated until the twenty-four columns are scanned, producing the desired forty-eight element vector for this character.

As stated earlier the reason for positioning the characters in the lower left corner of the matrix is to provide an accurate means of registration. The results of the classification process would be changed very little if the characters were encoded with the center of the line of type in the center of the matrix, i.e., element twenty-five of the vector corresponding to the bottom of the line of type instead of the bottom of the character. However, vertical registration is not necessarily accurately maintained on a typewriter, therefore, there is no assurance that the same character will always lie in the same vertical position on a line of type. Accurate registration under these conditions is very difficult. Even slight misregistration could alter the encoded representation of the character enough to cause it to be incorrectly classified or to be rejected entirely.

In the scanning procedure as described it is possible to confuse small ink specks or paper blemishes with the extremities of characters. For this reason a suitable checking procedure should be used to eliminate the selection of these small specks as character extremities. A suitable check can be made by requiring that at least three consecutive scans in the raster pattern must contain black area before the procedure recognizes an extremity. The first of these scans would then be recorded as the beginning of the character. If three consecutive scans were not encountered with black area, the black area found in them would be ignored. This procedure should apply for both line and character

finding scanning procedures. A procedure such as this would not preclude the identification of a character as small as a period since the period is four rows high and seven columns wide.

The Classification Procedure

The principal objective of this study, as stated earlier, is the evaluation of a procedure for the classification of 61 characters from the English alphabet. The procedure is one in which the vector representation for an unclassified character (as described above) is compared against a set of tabulated character vectors which have been stored in a digital memory. Each character in the set is represented by two stored vectors in the memory. One of the vectors gives the upper limit of the vector element values and the other gives the lower limit. These boundaries allow the procedure to handle characters which may be slightly distorted or incorrectly registered.

For purposes of this study only one sample of each letter and numeral was used in the analysis. The characters were typed on bond paper with a carbon ribbon in order to minimize distortion. In the hand transcription of these characters obvious defects due to paper fibers and paper blemishes were disregarded. The purpose of this procedure was to make all character perimeters smooth and to eliminate any defects which did not appear to be an essential part of the character. The character vectors obtained in this manner were termed the "ideal character vectors."

The ideal character vectors could never be expected to occur in practice. However, the vectors which are obtained should have values

which are distributed about these ideal vectors in some manner. Distortion and slight misregistration would cause increases or decreases in the counts of the vector elements. Thus, the distribution of the actual vectors about the ideal can be approximated by assuming that the upper and lower limits which were assigned in this study were obtained by adding and subtracting a distortion constant from the ideal vector element values, i.e., the same constant was subtracted from and added to each element of every vector.

If corresponding elements from two different characters in the set have a region between their boundaries which is common to both, then little information will be obtained from the use of this element in the vector as a means for distinguishing between these two characters. Conversely, if corresponding elements from two different characters do not have any values in common, then that element will be quite useful in distinguishing between these two characters. It is quite unlikely that in a character set of the type studied here there will ever be a single element that will not have a region of overlapping values, assuming a reasonable spread between the upper and lower limits, in common with at least one other character in the set. Thus, it requires several vector elements to properly classify any given character.

A program was written for the Burroughs 220 digital computer to evaluate the effectiveness of the proposed classification procedure. This program is listed in the appendix. The program begins by reading from IBM cards the element values for the ideal characters. With the character vectors stored in memory the program selects each vector in

turn and computes its upper and lower bounds. The boundaries are computed using twice the value of the distortion constant in order that they may be used to represent the boundaries of both the ideal character with which it will be compared and its own boundaries. Thus, if an element value of the character vector with which it is being compared lies between the computed upper and lower bounds of the corresponding element of the selected character then these two elements do have a region of overlapping values. The character vector selected is compared against every other character in the set including itself. Each element of the ideal character vectors which do not lie between the corresponding boundaries of the selected character vector are noted and the element number (1 through 48) is printed out.

The results of this program indicate those characters which are distinguishable from the selected character for each value of distortion considered. The selected character and an ideal character are assumed to be distinguishable if at least one element in the ideal character vector is not within the limits specified for the same element of the selected character. Thus, the only character in the set that should not have any element values listed in the printout would be the selected character itself, for if any other character appears without at least one element listed then that character would be indistinguishable from the selected character.

The character vector can also be thought of as defining a point in N space. The upper and lower boundaries for the vector define a solid in this N space. Two characters are indistinguishable, in this sense, when their solids overlap in all dimensions. The program which

evaluated the classification procedure produced the solid for each selected character by doubling the distortion constant. Thus, if the boundary for the selected character enclosed the point defined for an ideal character then the two characters were indistinguishable.

The treatment of distortion, as presented here, is not a realistic approximation to actual conditions. However, the method does give an indication as to the ability of the classification procedure to cope with distorted characters. Perhaps a better approximation to true distortion would be some sort of a statistical model such that the probability that a particular square in the matrix is black or white is given for each square in the matrix. A model such as this has its own disadvantages, such as the accuracy with which these probabilities can be specified for all degrees of distortion. These probabilities are usually obtained by acquiring a large number of samples of each character and computing the desired probabilities from this sample space. It was not felt that a procedure of this complexity was required in this phase of the evaluation of this classification procedure.

Before any character recognition procedure is implemented some means of evaluating the procedure with an accurate measure of distortion should be employed. Such an analysis is necessary only after the original hypotheses have been shown to be reasonably operational.

CHAPTER III

RESULTS

General

The evaluation of the classification procedure has shown that each of the 61 characters studied can be correctly distinguished from any of the other characters in the set when the "ideal character vectors" are compared without distortion. The inclusion of distortion through the distortion constant, discussed in the previous chapter, provided results which indicate that at least a small amount of distortion can be tolerated by the classification procedure.

Evaluation Program

The characters used by the evaluation program are shown in Figure 4. This figure shows each character correctly registered within a rectangular matrix. From this figure the features of a character located between specific rows or columns can be determined.

The evaluated procedure compared each of the ideal character vectors, after modification for distortion, against each ideal character vector (unmodified) in the set including itself. Each unmodified ideal vector element which did not lie between the limits defined by the corresponding modified character vector element was printed out and labeled with the name of the character from which it was obtained. These elements are the ones which can be used to distinguish between the two characters in question. Under ideal conditions only the comparison

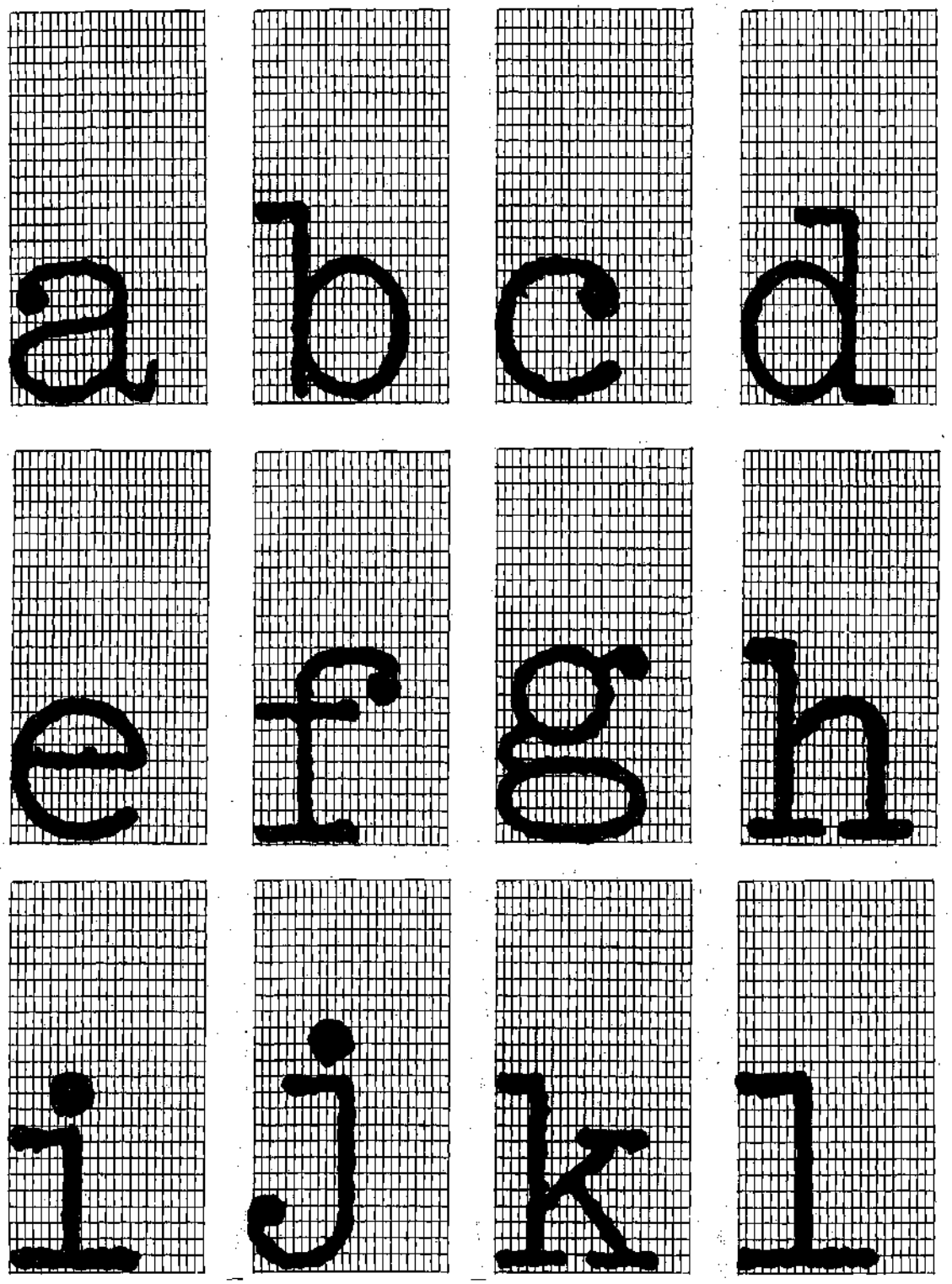


Figure. 4. The Character Set Correctly Registered in the Rectangular Matrices

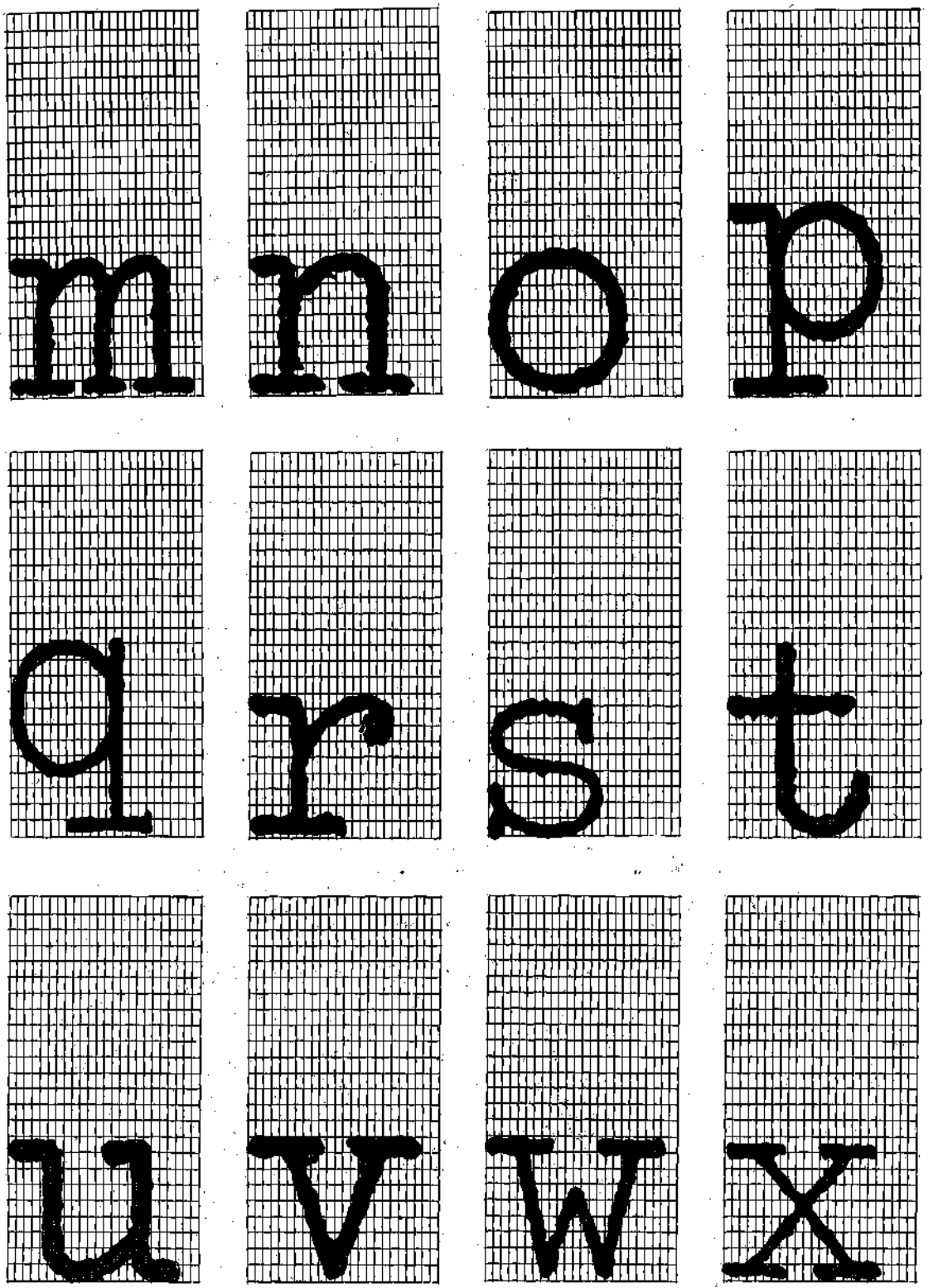


Figure 4. (Continued)

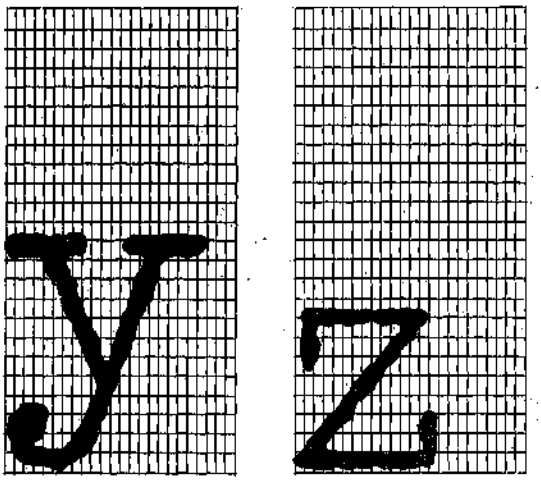


Figure 4. (Continued)

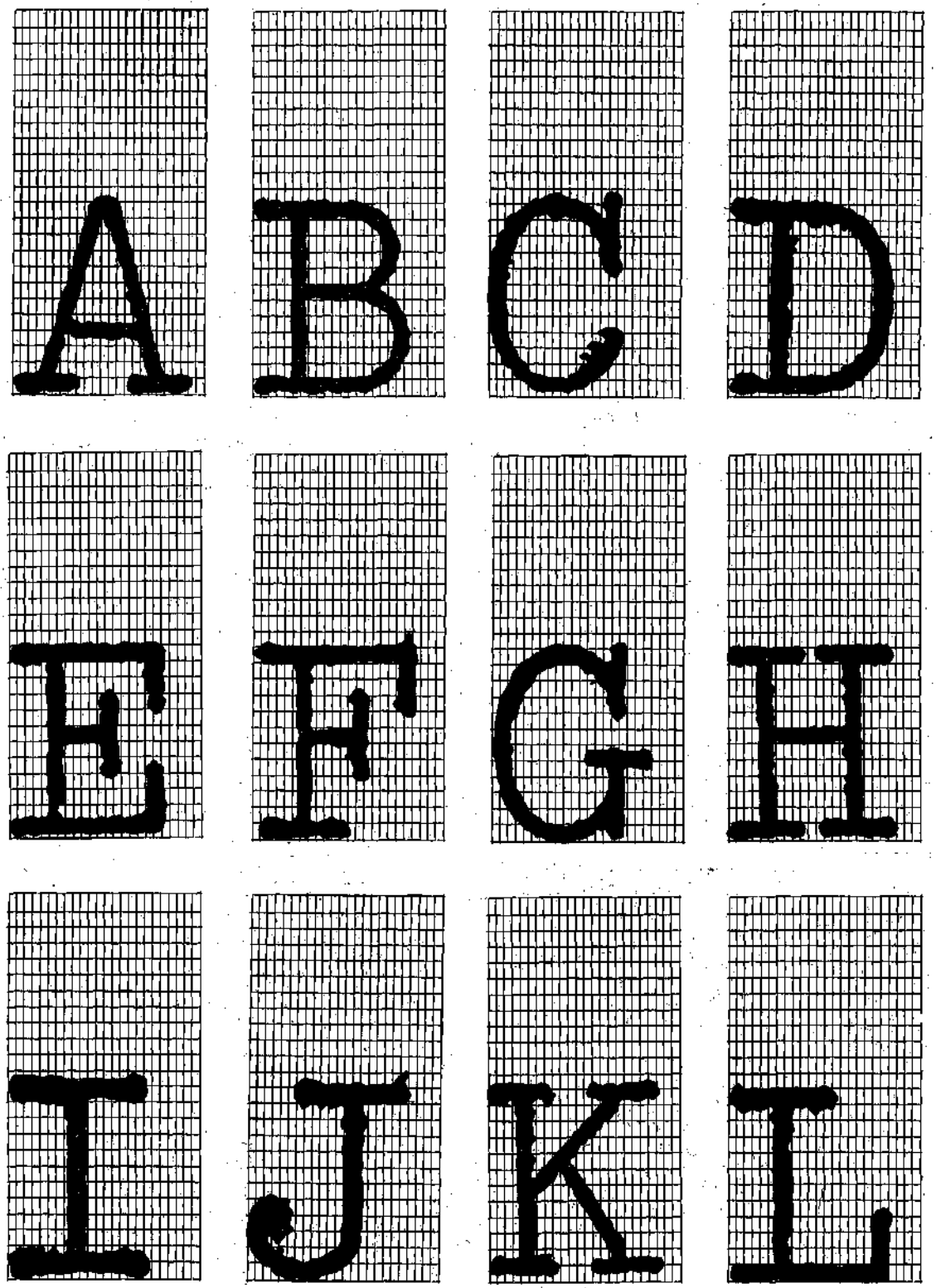


Figure 4. (Continued)

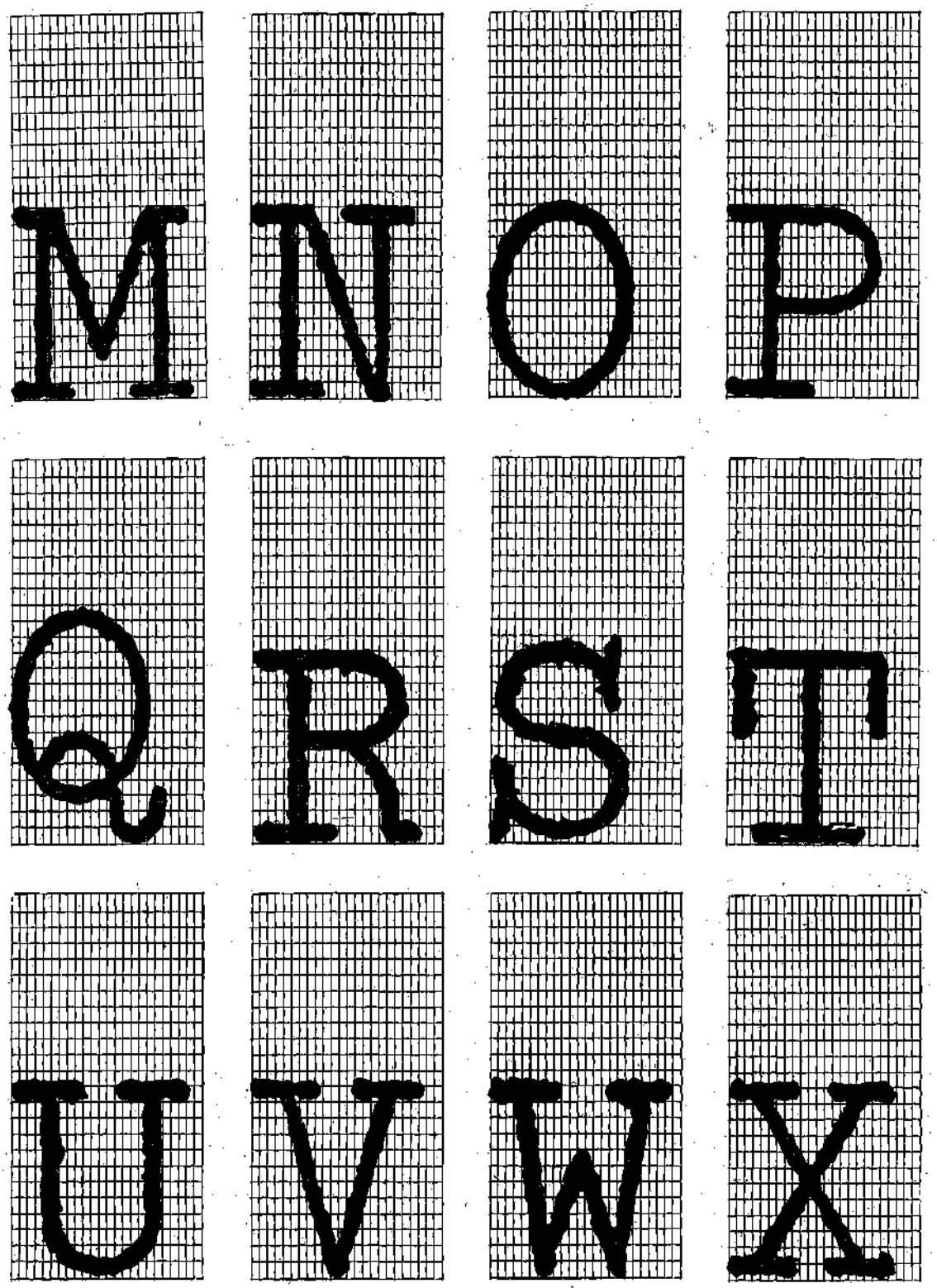


Figure 4. (Continued)

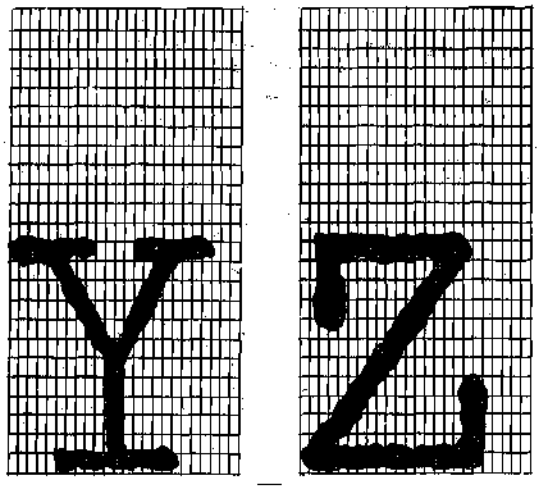


Figure 4. (Continued)

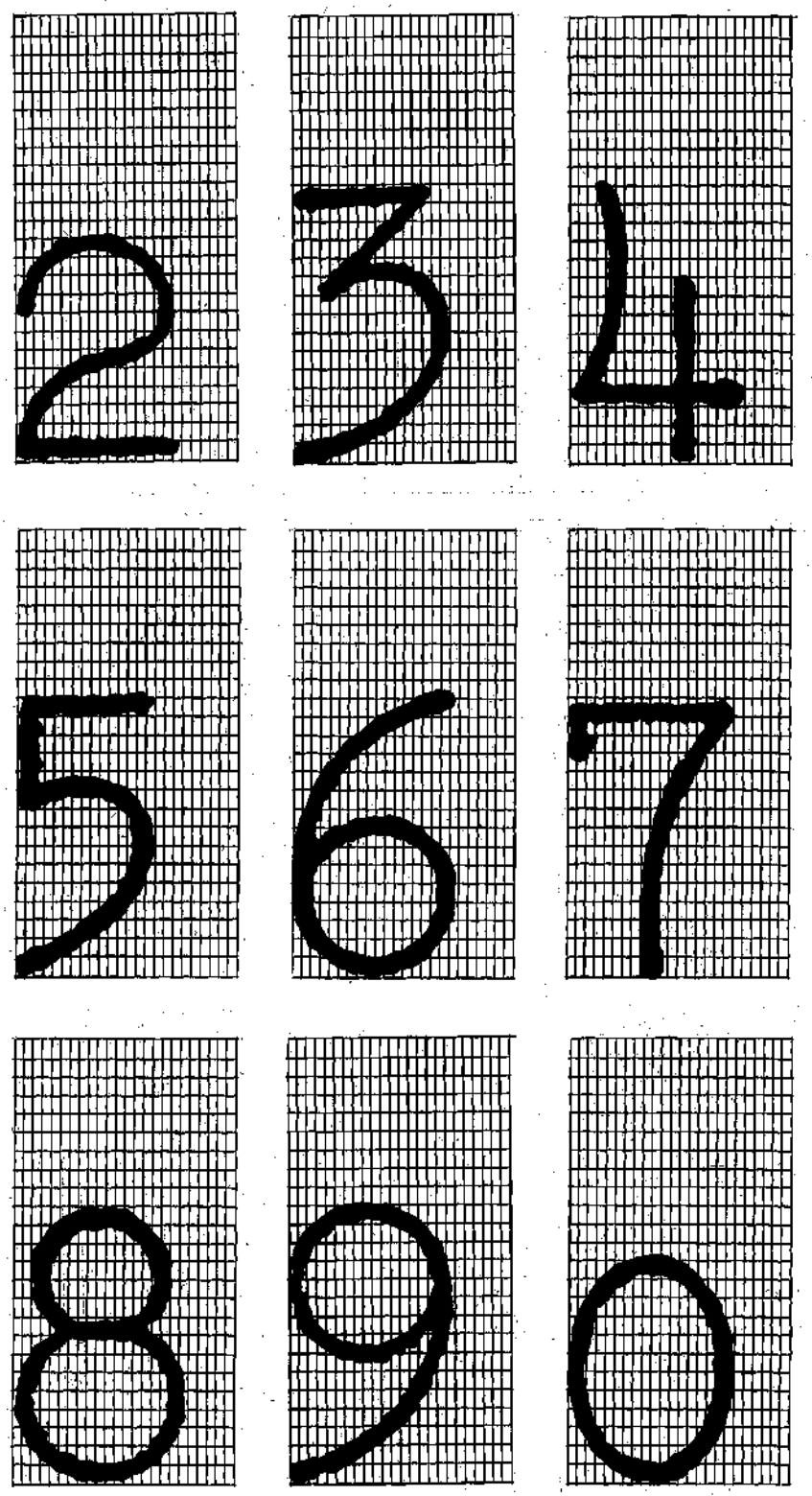


Figure 4. (Contimued)

between the distorted ideal vector and its undistorted counterpart in the set should have no elements printed out as being outside the limits specified by the distorted vector boundaries.

The results of this analysis showed that the row elements predominate as a means for distinguishing between two character vectors. Although the row elements were predominant the following nine character pairs could not be distinguished when the value of the distortion constant was set at two if only the row elements are used: b-d, u-x, u-z, B-E, B-H, D-X, D-Z, N-X, and X-Z. This means that in these nine cases every row element was found to be between the boundaries specified by the paired-distorted character, i.e., the row elements in the ideal undistorted z vector were within the boundaries set by the distorted u vector and conversely. Table 2 gives a list of the characters which were found to have fewer than five elements which could be used for purposes of distinguishing each character listed from the character paired with it in the list. The distinguishing elements are given beside the character pair.

Character Features

The encoding process provides a method for measuring the black area contained in a row or column of the rectangular encoding matrix. A distorted character will have an area increase or decrease in those rows or columns which intersect the distorted portions (distorted relative to the ideal character) of the character. It would be possible for a portion of a line in a distorted character to be entirely eliminated or for a speck to appear which would look like an additional line

Table 2. Character Comparison Table

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
a	a	0	0	0	0
a	c	1	2	5	10
a	d	2	5	8	12
a	e	2	3	5	9
a	o	2	5	38	39
a	s	2	39	0	0
a	x	1	2	5	8
a	z	2	5	8	39
b	b	0	0	0	0
b	d	30	31	38	39
b	f	9	0	0	0
b	f	9	0	0	0
b	h	1	29	31	40
b	k	1	29	31	0
b	p	4	8	11	12
b	r	12	0	0	0
b	D	1	8	11	12
b	O	8	26	30	31
c	a	1	2	5	10
c	c	0	0	0	0
c	o	2	10	41	0
c	s	1	5	9	10
c	v	2	8	10	27
c	w	2	4	10	27
c	x	1	2	8	10
c	y	7	9	12	13
c	6	15	40	41	0
d	a	2	5	8	12
d	b	30	31	38	39
d	d	0	0	0	0
d	k	1	29	30	38
d	q	5	8	13	39
d	s	5	12	38	39
d	x	9	12	38	0
d	z	9	12	38	39
d	A	5	8	12	38
d	J	8	11	12	37
d	2	7	8	38	39
e	a	2	3	5	9
e	e	0	0	0	0
e	o	5	8	0	0

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
e	r	5	7	8	30
e	s	5	37	0	0
e	v	1	5	8	9
e	x	1	5	9	0
e	y	5	8	12	13
e	A	1	5	8	10
e	O	5	8	12	41
e	9	5	10	15	41
f	b	9	0	0	0
f	f	0	0	0	0
f	h	7	9	29	31
f	k	9	12	29	31
g	g	0	0	0	0
g	O	6	8	12	0
g	S	6	7	26	42
h	b	1	29	31	40
h	f	7	9	29	31
h	h	0	0	0	0
h	k	0	0	0	0
h	n	12	30	0	0
h	u	9	12	0	0
h	x	9	12	29	30
h	A	5	8	29	30
h	K	8	11	12	28
h	2	7	8	29	30
i	i	0	0	0	0
i	l	8	9	0	0
i	A	5	32	33	0
i	2	6	32	33	0
j	j	0	0	0	0
k	b	1	29	31	0
k	d	1	29	30	38
k	f	9	12	29	31
k	h	0	0	0	0
k	k	0	0	0	0
k	n	12	0	0	0
k	x	8	12	29	30
k	z	12	29	30	0
k	K	11	12	28	0
k	2	8	9	29	30

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
k	9	1	15	29	30
l	i	8	9	0	0
l	l	0	0	0	0
l	t	1	2	9	12
l	A	5	12	32	33
l	I	2	12	13	34
l	2	6	32	33	0
m	m	0	0	0	0
m	z	43	44	0	0
n	h	12	30	0	0
n	k	12	0	0	0
n	n	0	0	0	0
n	o	1	8	29	30
n	r	1	29	31	0
n	u	1	0	0	0
n	v	1	9	29	41
n	w	1	8	9	29
n	x	9	29	0	0
n	z	9	40	41	0
n	A	5	8	10	29
n	2	8	11	12	0
n	9	1	10	15	29
o	a	2	5	38	39
o	c	2	10	41	0
o	e	5	8	0	0
o	n	1	8	29	30
o	o	0	0	0	0
o	t	9	32	33	0
o	u	1	8	29	30
o	v	8	9	0	0
o	w	1	9	0	0
o	x	1	8	9	0
o	y	9	12	13	0
o	A	1	5	9	10
o	C	10	11	12	0
o	V	9	10	11	12
o	X	1	10	11	12
o	0	9	12	0	0
o	2	1	11	12	0
o	3	1	9	11	15
o	6	15	0	0	0

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
o	8	8	13	14	15
o	9	8	10	15	0
p	b	4	8	11	12
p	p	0	0	0	0
p	D	4	0	0	0
p	L	1	4	32	0
q	d	5	8	13	39
q	q	0	0	0	0
q	0	2	27	38	41
q	0	5	13	38	41
q	2	1	13	38	0
r	b	12	0	0	0
r	e	5	7	8	30
r	n	1	29	31	0
r	r	0	0	0	0
r	v	1	7	9	30
r	x	9	37	0	0
r	z	9	0	0	0
s	a	2	39	0	0
s	c	1	5	9	10
s	d	5	12	38	39
s	e	5	37	0	0
s	s	0	0	0	0
s	x	5	8	9	0
s	z	5	8	9	0
s	A	10	26	0	0
t	l	1	2	9	12
t	o	9	32	33	0
t	t	0	0	0	0
t	v	2	8	32	33
t	w	2	5	32	33
u	h	9	12	0	0
u	n	1	0	0	0
u	o	1	8	29	30
u	u	0	0	0	0
u	x	40	41	0	0
u	z	40	41	0	0
u	N	10	11	12	42
v	c	2	8	10	27

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
v	e	1	5	8	9
v	n	1	9	29	41
v	o	8	9	0	0
v	r	1	7	9	30
v	t	2	8	32	33
v	v	0	0	0	0
v	w	8	0	0	0
v	x	1	0	0	0
v	Z	1	0	0	0
v	3	8	9	11	15
v	9	9	10	15	41
w	c	2	4	10	27
w	n	1	8	9	29
w	o	1	9	0	0
w	t	2	5	32	33
w	v	8	0	0	0
w	w	0	0	0	0
w	x	1	5	8	0
w	A	1	9	10	0
w	V	9	10	11	12
w	0	1	9	12	26
x	a	1	2	5	8
x	c	1	2	8	10
x	d	9	12	38	0
x	e	1	5	9	0
x	h	9	12	29	30
x	k	8	12	29	30
x	n	9	29	0	0
x	o	1	8	9	0
x	r	9	37	0	0
x	s	5	8	9	0
x	u	40	41	0	0
x	v	1	0	0	0
x	w	1	5	8	0
x	x	0	0	0	0
x	z	0	0	0	0
x	A	5	8	9	10
x	2	8	9	11	12
x	8	1	13	14	15
x	9	1	9	10	15
y	e	7	9	12	13

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
y	e	5	8	12	13
y	o	9	12	13	0
y	y	0	0	0	0
y	0	12	13	41	0
y	T	1	2	34	35
y	V	2	11	13	0
y	X	1	2	11	13
y	Y	1	35	36	0
y	0	2	12	13	41
y	3	11	12	13	15
z	a	2	5	8	39
z	c	1	2	8	10
z	d	9	12	38	39
z	e	1	5	9	0
z	k	12	29	30	0
z	m	43	44	0	0
z	n	9	40	41	0
z	r	9	0	0	0
z	s	5	8	9	0
z	u	40	41	0	0
z	v	1	0	0	0
z	x	0	0	0	0
z	z	0	0	0	0
A	d	5	8	12	38
A	e	1	5	8	10
A	h	5	8	29	30
A	i	5	32	33	0
A	l	5	12	32	33
A	n	5	8	10	29
A	o	1	5	9	10
A	s	10	26	0	0
A	w	1	9	10	0
A	x	5	8	9	10
A	A	0	0	0	0
A	V	1	5	11	12
A	X	5	11	12	0
A	Z	11	12	27	28
A	0	1	5	26	0
A	2	5	12	26	27
A	3	1	5	11	15
B	B	0	0	0	0

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
B	D	6	7	41	0
B	E	41	42	44	0
B	F	1	41	0	0
B	H	29	31	40	42
B	P	1	7	29	41
B	R	6	11	29	31
B	W	1	0	0	0
C	o	10	11	12	0
C	C	0	0	0	0
C	G	5	6	0	0
C	S	2	7	8	13
C	V	1	12	25	26
C	X	1	12	26	0
C	O	11	39	0	0
C	2	1	6	0	0
C	3	1	12	15	26
C	9	8	11	12	15
D	b	1	8	11	12
D	p	4	0	0	0
D	B	6	7	41	0
D	D	0	0	0	0
D	L	11	32	42	0
D	P	6	29	0	0
D	W	1	4	6	7
D	X	30	31	0	0
D	Z	28	29	30	31
E	B	41	42	44	0
E	E	0	0	0	0
E	F	1	0	0	0
F	B	1	41	0	0
F	E	1	0	0	0
F	F	0	0	0	0
F	W	2	3	4	32
G	C	5	6	0	0
G	G	0	0	0	0
G	J	6	11	37	40
G	2	1	5	39	40
H	B	29	31	40	42
H	H	0	0	0	0
H	R	1	6	11	40

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
I	l	2	12	13	34
I	I	0	0	0	0
J	d	8	11	12	37
J	G	6	11	37	40
J	J	0	0	0	0
J	V	1	37	38	39
J	X	1	30	37	38
K	h	8	11	12	28
K	k	11	12	28	0
K	K	0	0	0	0
K	N	1	28	41	42
K	R	7	28	41	42
K	V	1	28	29	30
K	X	7	28	29	30
L	p	1	4	32	0
L	D	11	32	42	0
L	L	0	0	0	0
M	M	0	0	0	0
M	Z	6	7	43	44
N	u	10	11	12	42
N	K	1	28	41	42
N	N	0	0	0	0
N	R	7	42	0	0
N	X	29	30	41	42
O	g	6	8	12	0
O	q	2	27	38	41
O	y	12	13	41	0
O	O	0	0	0	0
O	V	2	12	13	27
O	0	2	13	27	0
O	2	1	2	13	42
P	B	1	7	29	41
P	D	6	29	0	0
P	P	0	0	0	0
Q	Q	0	0	0	0
Q	9	4	6	7	8
R	B	6	11	29	31
R	H	1	6	11	40

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
R	K	7	28	41	42
R	N	7	42	0	0
R	R	0	0	0	0
R	V	1	7	29	30
R	X	7	29	30	41
S	g	6	7	26	42
S	C	2	7	8	13
S	S	0	0	0	0
T	Y	1	2	34	35
T	T	0	0	0	0
T	X	2	13	34	35
T	Y	12	13	34	35
U	U	0	0	0	0
V	o	9	10	11	12
V	W	9	10	11	12
V	y	2	11	13	0
V	A	1	5	11	12
V	C	1	12	25	26
V	J	1	37	38	39
V	K	1	28	29	30
V	O	2	12	13	27
V	R	1	7	29	30
V	V	0	0	0	0
V	X	1	0	0	0
V	0	11	12	26	0
V	2	1	11	12	0
V	3	12	15	0	0
V	9	8	11	12	15
W	B	1	0	0	0
W	D	1	4	6	7
W	F	2	3	4	32
W	W	0	0	0	0
X	o	1	10	11	12
X	y	1	2	11	13
X	A	5	11	12	0
X	C	1	12	26	0
X	D	30	31	0	0
X	J	1	30	37	38
X	K	7	28	29	30
X	N	29	30	41	42

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
X	R	7	29	30	41
X	T	2	13	34	35
X	V	1	0	0	0
X	X	0	0	0	0
X	Z	28	0	0	0
X	0	1	11	12	26
X	2	11	12	0	0
X	3	1	12	15	0
Y	y	1	35	36	0
Y	T	12	13	34	35
Y	Y	0	0	0	0
Z	A	11	12	27	28
Z	D	28	29	30	31
Z	M	6	7	43	44
Z	X	28	0	0	0
Z	Z	0	0	0	0
0	b	8	26	30	31
0	e	5	8	12	41
0	o	9	12	0	0
0	q	5	13	38	41
0	w	1	9	12	26
0	y	2	12	13	41
0	A	1	5	26	0
0	C	11	39	0	0
0	O	2	13	27	0
0	V	11	12	26	0
0	X	1	11	12	26
0	0	0	0	0	0
0	2	1	0	0	0
0	5	10	12	15	41
0	9	8	15	0	0
2	d	7	8	38	39
2	h	7	8	29	30
2	i	6	32	33	0
2	k	8	9	29	30
2	l	6	32	33	0
2	n	8	11	12	0
2	o	1	11	12	0
2	q	1	13	38	0
2	x	8	9	11	12
2	A	5	12	26	27

(Continued)

Table 2. (Continued)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
2	C	1	6	0	0
2	G	1	5	39	40
2	O	1	2	13	42
2	V	1	11	12	0
2	X	11	12	0	0
2	0	1	0	0	0
2	2	0	0	0	0
2	9	1	8	12	15
3	o	1	9	11	15
3	v	8	9	11	15
3	y	11	12	13	15
3	A	1	5	11	15
3	C	1	12	15	26
3	V	12	15	0	0
3	X	1	12	15	0
3	3	0	0	0	0
3	5	10	25	26	40
3	6	8	9	11	15
3	7	2	11	34	0
3	9	2	8	11	15
4	4	0	0	0	0
5	O	10	12	15	41
5	3	10	25	26	40
5	5	0	0	0	0
6	c	15	40	41	0
6	o	15	0	0	0
6	3	8	9	11	15
6	6	0	0	0	0
6	8	2	14	0	0
6	9	2	8	9	0
7	3	2	11	34	0
7	7	0	0	0	0
8	o	8	13	14	15
8	x	1	13	14	15
8	6	2	14	0	0
8	8	0	0	0	0
8	9	9	14	0	0
9	e	5	10	15	41

(Continued)

Table 2. (Concluded)

Distorted Character	Ideal Character	Element Numbers of Ideal Character Which Do Not Lie Within Limits Defined by the Same Distorted Character Element			
9	k	1	15	29	30
9	n	1	10	15	29
9	o	8	10	15	0
9	v	9	10	15	41
9	x	1	9	10	15
9	C	8	11	12	15
9	Q	4	6	7	8
9	V	8	11	12	15
9	0	8	15	0	0
9	2	1	8	12	15
9	3	2	8	11	15
9	6	2	8	9	0
9	8	9	14	0	0
9	9	0	0	0	0

segment.

If the area determined for a particular ideal vector element is relatively high any changes from character sample to character sample, brought about by distortion, will represent a much smaller percentage change in the feature involved than if the feature were small (low area count.) Thus, distortion would cause a relatively high percentage change. This implies that in order to consider distortion in the classification procedure, vector elements which have relatively low counts, with respect to the allowable distortion, should be discarded and elements which contain counts which are very close to a corresponding element in another character should be ignored in the process of distinguishing between these two characters. A reliable classification procedure should be based upon those elements which would not be expected to have their area count vanish due to the effects of distortion.

The elements which possess relatively high counts can be obtained from either or both of two situations in a character. First, the associated row or column may pass along, or parallel to, a long straight segment or an extremity of a long slow curve. Or second, the associated row or column could be crossed several times by character segments. In either case a relatively high area count would be produced.

Curved segments can produce high area counts only at their left, right, top, or bottom extremities since the encoding process is horizontally and vertically oriented. Straight segments produce high area counts only when they are horizontal or vertical. Slanted segments and certain portions of curves look alike in either the rows or columns

and produce approximately the same area count as a single perpendicular crossing of the row or column. Those segments which produce low area counts are important when several of them cross the same row or column, as is the case in an m and w for the row elements which pass near the center of these characters. Thus, the encoding process tends to cause the characters to take on a block format.

The effect of vector elements which have low area counts and also of elements which have counts relatively close to the count of a corresponding element in another character is illustrated in Table 3. Three increments of the distortion constant are shown in this table, zero, one, and two. When the distortion constant is zero the table shows a shift to the left indicating that only a few elements are indistinguishable in each vector. A constant of one produces a noticeable shift to the right indicating that there are numerous elements with very low values or which are very close to the values of other elements in different vectors. (This can also be verified from the histograms in Figure 5 which plot area count against number of elements possessing that count.) Again a marked shift to the right is evident when the distortion constant is increased to two. This value of distortion results in the inability to distinguish the lower case h from the lower case k and also the lower case x from the lower case z. All of the upper case and numeral characters are still distinguishable. Although, as can be seen in Table 3, many of the letters appear to have very few differences.

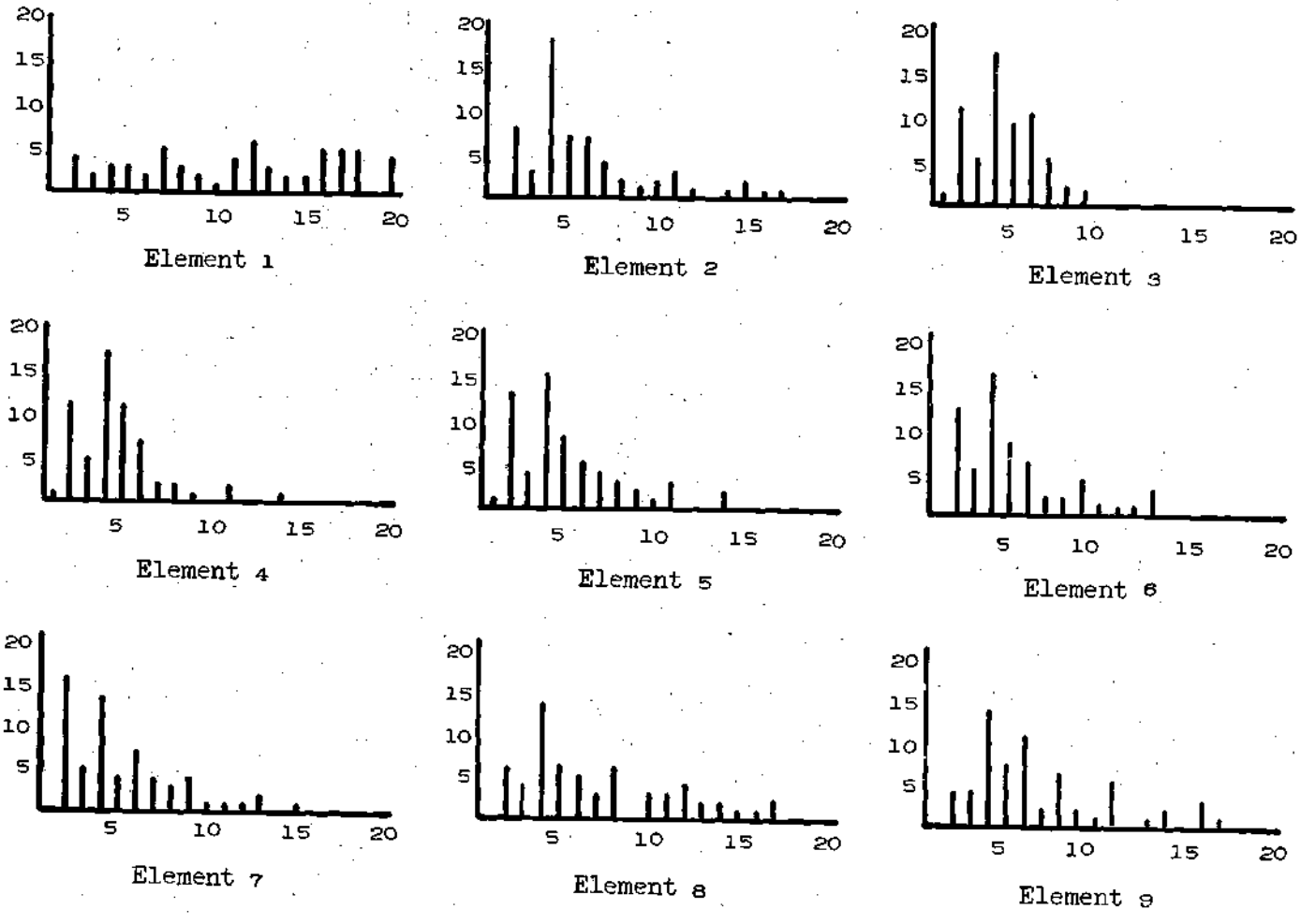
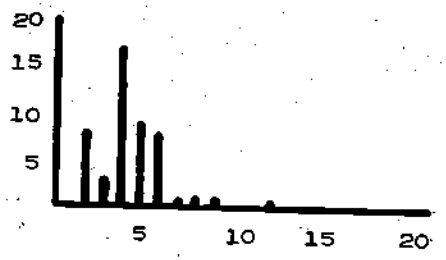
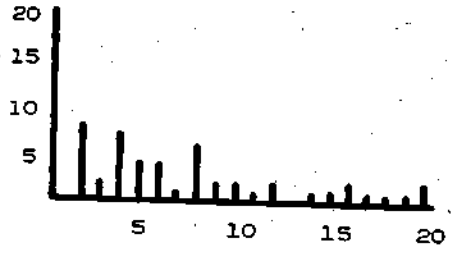


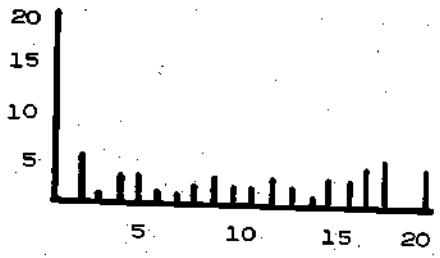
Figure 5. The Distribution of Values for Each Vector Element



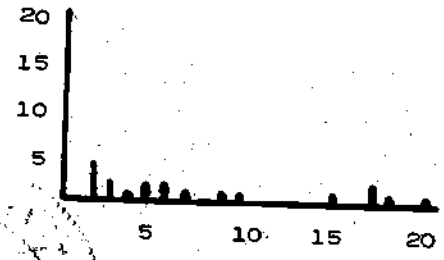
Element 10



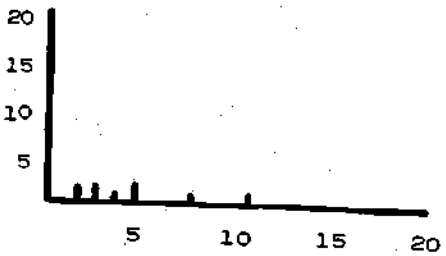
Element 11



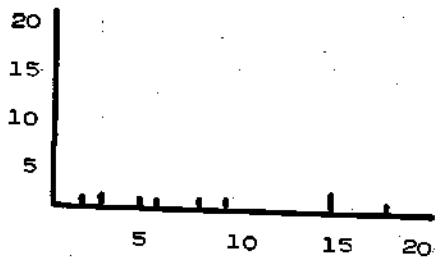
Element 12



Element 13



Element 14



Element 15

Figure 5. (Continued)

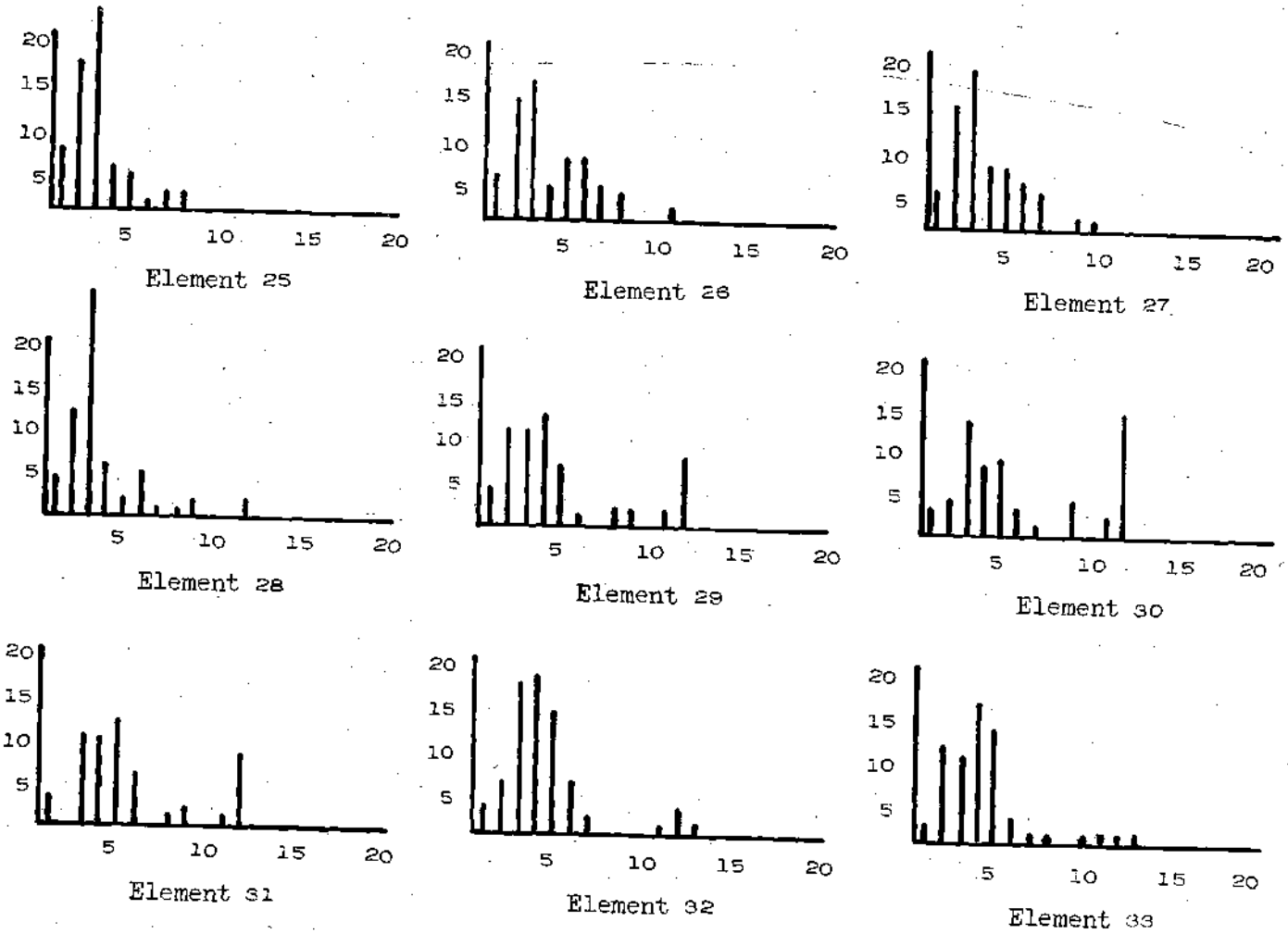


Figure 5. (Continued)

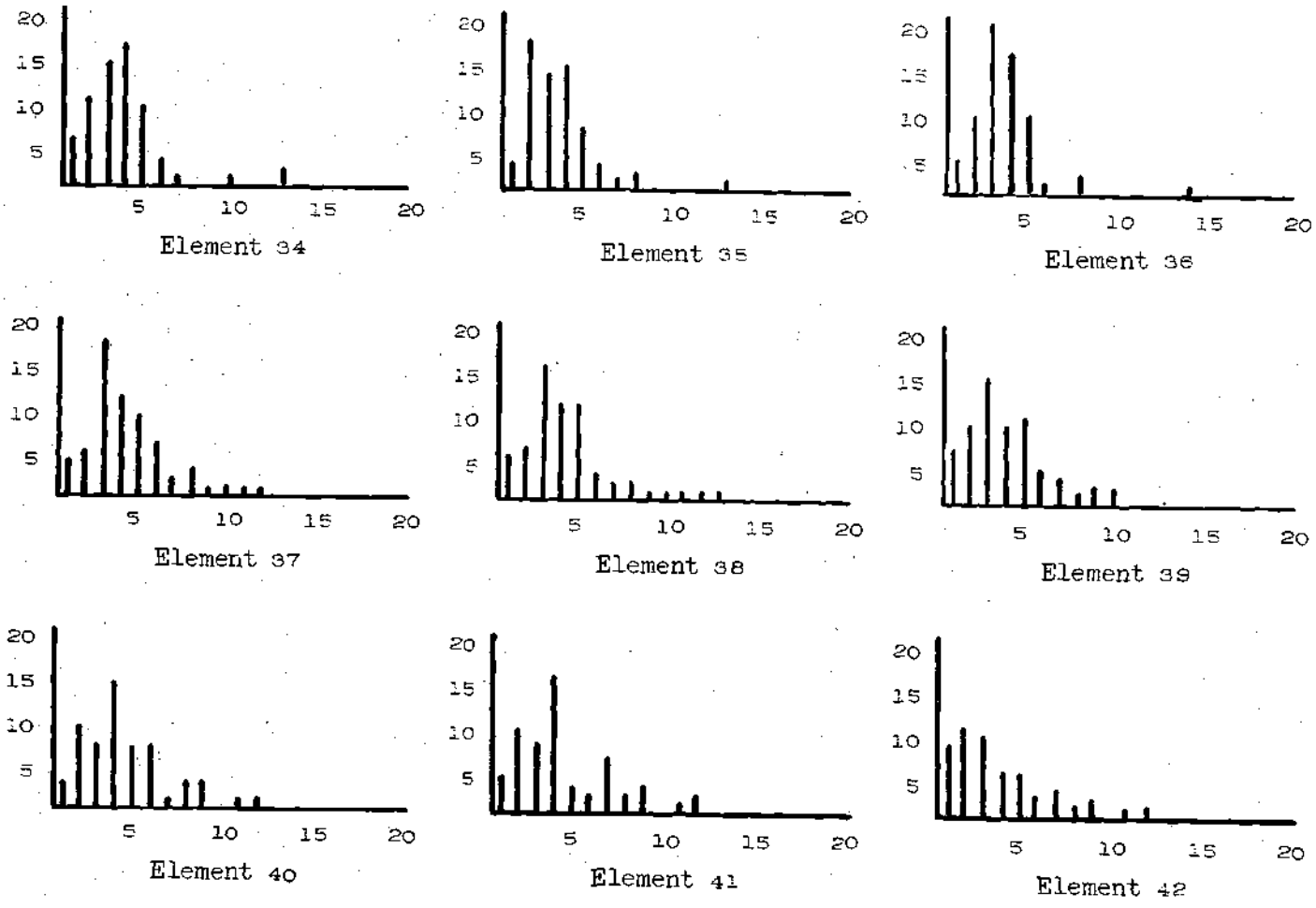


Figure 5. (Continued)

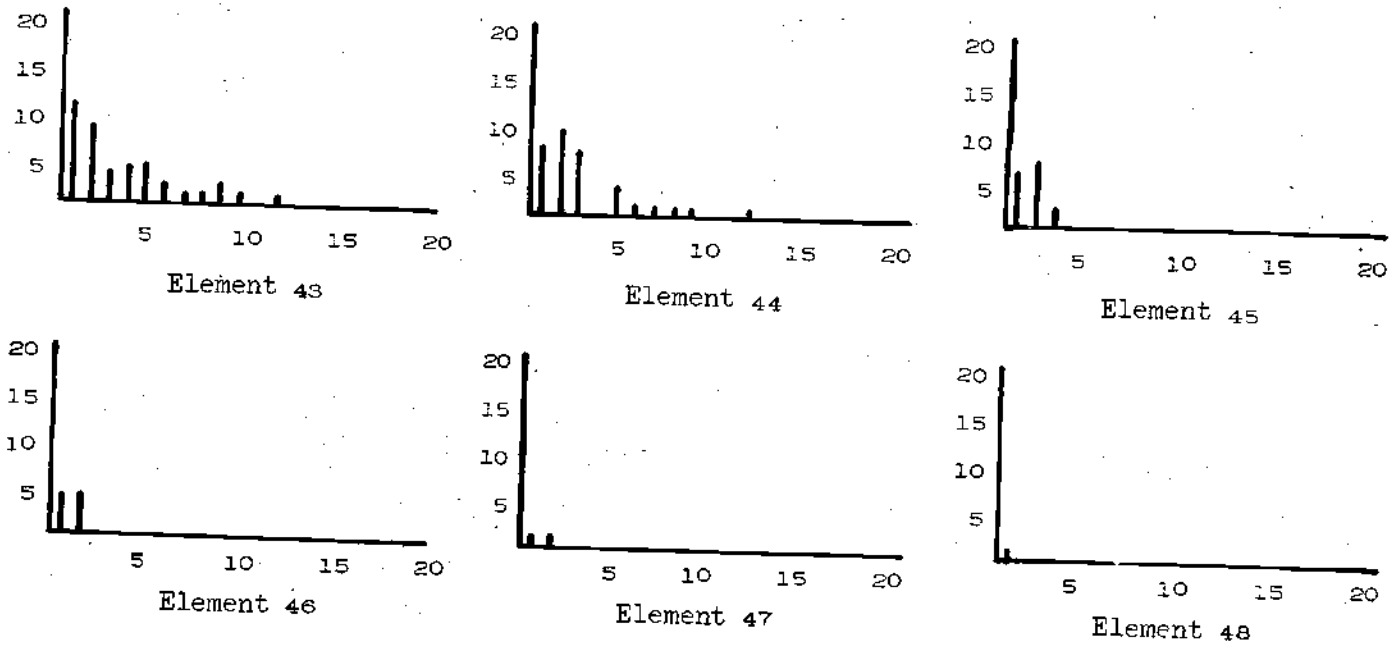


Figure 5. (Continued)

Table 3. A Comparison Between the Three Values of the Distortion Constant

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	42	17	1	0	0	0	0	1
a	1	0	0	13	29	13	4	1	1
	2	0	0	0	0	13	26	20	2
	0	41	15	3	1	0	0	0	1
b	1	0	0	4	17	20	15	4	1
	2	0	0	0	0	6	22	29	4
	0	40	14	5	1	0	0	0	1
c	1	0	0	1	27	23	8	1	1
	2	0	0	0	0	6	29	23	3
	0	38	19	3	0	0	0	0	1
d	1	0	0	1	15	27	14	3	1
	2	0	0	0	0	4	26	29	2
	0	29	21	9	1	0	0	0	1
e	1	0	0	1	16	20	19	4	1
	2	0	0	0	0	8	17	31	5
	0	29	23	7	1	0	0	0	1
f	1	0	0	3	15	23	19	0	1
	2	0	0	0	0	8	28	23	2
	0	52	8	0	0	0	0	0	1
g	1	0	6	16	24	11	3	0	1
	2	0	0	0	2	8	37	12	2

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	40	11	6	3	0	0	0	1
h	1	0	0	4	17	21	13	5	1
	2	0	0	0	0	5	29	23	4
	0	35	20	2	3	0	0	0	1
i	1	0	3	10	19	25	2	1	1
	2	0	0	0	2	9	29	18	3
	0	52	8	0	0	0	0	0	1
j	1	0	6	20	24	10	0	0	1
	2	0	0	0	2	18	30	10	1
	0	43	15	2	0	0	0	0	1
k	1	0	0	7	14	25	11	3	1
	2	0	0	0	0	5	17	33	6
	0	39	16	4	1	0	0	0	1
l	1	2	4	15	15	12	10	2	1
	2	0	0	0	5	13	22	18	3
	0	53	7	0	0	0	0	0	1
m	1	1	12	18	21	8	0	0	1
	2	0	0	0	9	30	16	4	2
	0	33	16	10	1	0	0	0	1
n	1	0	0	6	16	22	12	4	1
	2	0	0	0	0	7	25	21	8

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	35	18	6	0	1	0	0	1
o	1	0	0	7	6	26	13	8	1
	2	0	0	0	0	3	22	23	13
	0	46	13	1	0	0	0	0	1
p	1	0	1	13	22	17	7	0	1
	2	0	0	0	1	9	33	15	3
	0	46	12	2	0	0	0	0	1
q	1	0	0	2	19	27	11	1	1
	2	0	0	0	0	8	25	26	2
	0	33	19	6	2	0	0	0	1
r	1	0	2	6	17	22	11	2	1
	2	0	0	0	0	7	27	22	5
	0	34	18	8	0	0	0	0	1
s	1	0	1	5	18	24	8	4	1
	2	0	0	0	0	6	24	25	6
	0	44	12	2	2	0	0	0	1
t	1	0	1	6	20	17	14	2	1
	2	0	0	0	4	12	27	16	2
	0	46	9	4	1	0	0	0	1
u	1	0	4	10	22	15	5	4	1
	2	0	0	0	2	12	23	19	5

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
v	0	33	20	7	0	0	0	0	1
	1	0	0	1	9	23	22	4	2
	2	0	0	0	0	5	24	27	5
w	0	50	8	2	0	0	0	0	1
	1	0	0	4	18	28	8	1	2
	2	0	0	0	0	5	30	21	5
x	0	36	17	6	1	0	0	0	1
	1	0	0	4	21	19	13	3	1
	2	0	0	0	0	1	27	22	11
y	0	52	7	1	0	0	0	0	1
	1	0	0	4	17	28	10	1	1
	2	0	0	0	1	2	17	36	5
z	0	40	16	4	0	0	0	0	1
	1	0	0	7	24	20	5	4	1
	2	0	0	0	0	4	25	22	10
A	0	46	10	4	0	0	0	0	1
	1	0	0	2	7	31	17	3	1
	2	0	0	0	0	3	19	36	3
B	0	45	9	6	0	0	0	0	1
	1	0	4	8	21	14	12	1	1
	2	0	0	0	1	19	25	11	5

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	39	19	2	0	0	0	0	1
C	1	0	0	8	24	22	5	1	1
	2	0	0	0	1	4	32	18	6
	0	31	15	14	0	0	0	0	1
D	1	0	0	3	13	22	17	5	1
	2	0	0	0	0	5	31	19	6
	0	42	13	3	2	0	0	0	1
E	1	0	0	12	27	14	6	0	2
	2	0	0	0	2	22	25	9	3
	0	42	13	4	1	0	0	0	1
F	1	0	1	13	26	14	4	1	2
	2	0	0	0	4	18	22	14	3
	0	44	16	0	0	0	0	0	1
G	1	0	0	6	28	18	8	0	1
	2	0	0	0	1	7	37	14	2
	0	39	13	6	2	0	0	0	1
H	1	0	1	14	18	19	7	1	1
	2	0	0	0	0	14	33	13	1
	0	47	12	1	0	0	0	0	1
I	1	0	1	6	15	24	12	2	1
	2	0	0	0	2	12	37	9	1

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	53	7	0	0	0	0	0	1
J	1	0	0	4	25	24	7	0	1
	2	0	0	0	0	12	30	18	1
	0	50	6	4	0	0	0	0	1
K	1	0	1	4	22	20	10	3	1
	2	0	0	0	0	5	34	20	2
	0	43	15	2	0	0	0	0	1
L	1	0	1	8	19	25	6	1	1
	2	0	0	0	2	10	32	14	3
	0	57	3	0	0	0	0	0	1
M	1	0	2	10	33	15	0	0	1
	2	0	0	1	8	25	22	4	1
	0	48	8	4	0	0	0	0	1
N	1	0	0	12	17	22	6	3	1
	2	0	0	0	1	8	35	15	2
	0	43	14	3	0	0	0	0	1
O	1	0	0	3	16	25	14	2	1
	2	0	0	0	0	4	35	18	4
	0	33	19	8	0	0	0	0	1
P	1	0	1	3	17	23	11	5	1
	2	0	0	0	0	8	32	19	2

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
Q	0	59	1	0	0	0	0	0	1
	1	2	2	15	27	13	1	0	1
	2	0	0	0	0	17	35	8	1
R	0	40	13	5	2	0	0	0	1
	1	0	1	3	12	22	15	7	1
	2	0	0	0	0	7	32	20	2
S	0	54	6	0	0	0	0	0	1
	1	1	7	22	20	8	2	0	1
	2	0	0	0	6	21	25	8	1
T	0	51	7	2	0	0	0	0	1
	1	0	5	17	28	10	0	0	1
	2	0	0	0	2	26	27	5	1
U	0	54	5	1	0	0	0	0	1
	1	0	2	7	27	17	6	1	1
	2	0	0	0	0	25	30	5	1
V	0	46	11	3	0	0	0	0	1
	1	0	1	0	5	25	22	6	2
	2	0	0	0	0	1	11	43	6
W	0	60	0	0	0	0	0	0	1
	1	0	2	20	26	9	3	0	1
	2	0	0	1	6	20	21	11	2

(Continued)

Table 3. (Continued)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	47	9	4	0	0	0	0	1
X	1	0	2	3	19	22	11	2	2
	2	0	0	0	0	1	14	39	7
	0	48	10	2	0	0	0	0	1
Y	1	0	0	9	24	19	8	0	1
	2	0	0	0	2	5	31	21	2
	0	48	12	0	0	0	0	0	1
Z	1	0	3	13	19	20	5	0	1
	2	0	0	0	0	10	27	22	2
	0	37	18	4	0	1	0	0	1
0	1	0	0	4	11	23	17	5	1
	2	0	0	0	0	4	11	39	7
	0	30	20	10	0	0	0	0	1
2	1	0	0	1	10	30	14	5	1
	2	0	0	0	0	2	13	36	10
	0	39	14	7	0	0	0	0	1
3	1	0	1	0	15	24	17	3	1
	2	0	0	0	0	6	21	30	4
	0	58	2	0	0	0	0	0	1
4	1	0	3	8	14	30	5	0	1
	2	0	0	0	3	9	38	10	1

(Continued)

Table 3. (Concluded)

Character	Distortion Constant	Number of Characters which have the Indicated Number of Indistinguishable Vector Elements for the Specified Value of the Distortion Constant							
		0 to 20	21 to 24	25 to 28	29 to 32	33 to 36	37 to 40	41 to 44	45 to 49
	0	45	14	1	0	0	0	0	1
5	1	0	1	12	22	20	4	1	1
	2	0	0	0	3	10	31	16	1
	0	53	7	0	0	0	0	0	1
6	1	0	1	11	25	19	4	0	1
	2	0	0	0	1	10	22	23	5
	0	47	9	3	1	0	0	0	1
7	1	0	9	12	14	20	5	0	1
	2	0	0	2	4	19	20	14	2
	0	47	12	1	0	0	0	0	1
8	1	0	1	12	20	22	4	1	1
	2	0	0	0	3	7	31	17	3
	0	52	7	1	0	0	0	0	1
9	1	0	1	9	18	17	12	3	1
	2	0	0	0	0	6	20	30	5

Orientation of Character Features

Figure 5 illustrates the distribution of the area counts for each of the vector elements. Vector elements 16 through 24, which are the top rows of the matrix, do not contain any black area from any of the characters in the set. If the slash mark or some of the special characters had been considered in the evaluation, these elements would have contained information. As pointed out previously, the row elements were predominant as a means for distinguishing between characters. This is also evident in Figure 5 as shown by the wide-flat distributions found in the row elements relative to the clustered patterns of low dispersion found in the column elements. The wide dispersion of the row elements obviously provides better classification because there are fewer elements with area counts which are relatively close together. The row elements also tend to have higher area counts than the columns. This would indicate that there are more horizontal features than vertical features. This can be born out by looking at the upper case letters which generally are in a block format. The vertical features tend to have at least one and often more horizontal segments protruding from them. In the lower case letters the features tend to be curved and thus it is more difficult to determine which features are horizontal or vertical. It would appear, though, that the longer segments in the lower case letters are vertical but it would be difficult to decide, excluding these long segments, whether the lower case letters have more horizontal or vertical features. The numbers are principally curves and slanted lines. The features which could be considered as horizontal or vertical seem to have little bearing upon the

classification of numbers which are quite similar since these tend to have very few straight horizontal or vertical features.

CHAPTER IV

CONCLUSIONS AND RECOMMENDATIONS

The results of the evaluation of this classification procedure indicate that reliable classification of the characters in the selected character set can be achieved even when some degree of distortion is considered. However, no evaluation was made which could determine the effect of slight vertical or horizontal misregistration within the encoding matrix. Misregistration might appear as slight translations caused by distortion in the left or lower extremity of the character. Therefore, it is not possible to establish the effectiveness of the procedure under realistic constraints until this information is available. These preliminary results indicate that some degree of freedom is probably available for considering translation

It is further recommended that a better analysis of distortion be undertaken before drawing any conclusions concerning the potential of this procedure. This analysis might present the procedure with numerous samples of each character which represent the entire range of expected distortion.

In summary, it would appear that this procedure does have certain advantages in simplicity over existing techniques. The results indicate that under realistic constraints the procedure would probably operate successfully with limited distortion permitted.

APPENDIX

COMPUTER PROGRAM FOR EVALUATION

OF THE CLASSIFICATION PROCEDURE

```
BAC-220 STANDARD VERSION  2/1/62
       INTEGER OTHERWISE  $
       ARRAY W(48)  $
       ARRAY T(61,48), CHAR(61), P(48), X(48)  $
       READ($$ DATA1)  $
       INPUT DATA1( SCAN, TOTAL, FOR I = ( 1,1,TOTAL )  $ CHAR(I),
              FOR I = (1,1,TOTAL)  $ FOR J = (1,1,SCAN)
              $ T(I,J) ) $
START..
       READ($$ DATA2 )  $
       INPUT DATA2( DELTA )  $
       FOR I = (1,1,TOTAL )  $
BEGIN
       FOR J = (1,1,SCAN)  $
BEGIN
       EITHER IF PCS(1)  $
BEGIN
       EITHER IF ( J EQL 1 ) OR ) J EQL 25 )  $
       X(J) = MAX( T(I,J), T(I,J+1) )  $
       OR IF ( J EQL SCAN ) OR ( J EQL 24 )  $
       X(J) = MAX( T(I,J-1), T(I,J) )  $
       OTHERWISE  $
```

```
        X(J) = MAX( T(I,J-1), T(I,J), T(I,J+1) 0  $
END  $
        OTHERWISE  $
        X(J) = T(I,J)  $
END  $
        FOR J = (1,1,SCAN)  $
BEGIN
        W(J) = X(J)  + 2.DELTA  $
        X(J) = X(J)  + 2.DELTA  $
END  $
        FOR K = (1,1,TOTAL)  $
BEGIN
        FOR J = (1,1,SCAN)  $
        P(J) = 0  $
        M = 0  $
        CH = CHAR(K)  $
        FOR J = (1,1,SCAN)  $
BEGIN
        EITHER IF PCS (2)  $
BEGIN
        EITHER IF ( J EQL 1 ) OR ( J EQL 25 )  $
        Y = MAX( T(K,J), T(K,J+1) )  $
        OR IF ( J EQL SCAN ) OR ( J EQL 24 )  $
        Y = MAX( T(K,J), T(K,J-1) )  $
        OTHERWISE  $
        Y = MAX( T(K,J-1), T(K,J), T(K,J+1) )  $
```

```
END  $
        OTHERWISE  $
        Y = T(K,J)  $
        IF ( W(J) LSS Y ) OR ( X(J) GTR Y )  $
BEGIN
        M = M+1  $
        P(M) = J  $
END  $
END  $
        WRITE($$ PRT, PT )  $
        OUTPUT PRT( DELTA, CHAR(I), CH, FOR J = (1,1,M)
                $ P(J) )  $
        FORMAT PT( I2, B2, A3, B2, A3, ( 24( B2, I2), WO,
                B12) )  $
END  $
END  $
        STOP  $
        GO TO START  $
FINISH  $
```

BIBLIOGRAPHY

Literature Cited

1. C. Heasly Clyde, Jr., and George L Fischer, Jr., "Some Elements of Optical Scanning," Optical Character Recognition, Spartan Books, 1962, pp. 15-27.

2. I. Flores and F. Ragonese, "Method of Synthesizing Waveform Generated by Character, Printed in Magnetic Ink, in Passing Beneath Magnetic Reading Head," Institute of Radio Engineers Transactions on Electronic Computers, Vol. EC-7, No. 4, December 1958, pp. 277-282.

3. E. C. Greanias, "Some Important Factors in the Practical Utilization of Optical Character Readers," Optical Character Recognition, Spartan Books, 1962, pp. 129-146.

4. "Electronic Retina Character Reader," Computer Design, April 1963, p. 3.

5. E. E. David, Jr., and O. G. Selfridge, "Eyes and Ears of Computers," Proceedings of the Institute of Radio Engineers, Vol. 50, No. 5, May 1962, pp. 1093-1101.

6. F. Rosenblatt, "Perceptron Simulation Experiments," Proceedings of the Institute of Radio Engineers, Vol. 48, No. 3, March 1960, pp. 301-309.

7. W. J. Hannan, "The RCA Multi-Font Reading Machine," Optical Character Recognition, Spartan Books, 1962, pp. 3-14.

8. L. Alt Franz, "Digital Pattern Recognition by Moments," Optical Character Recognition, Spartan Books, 1962, pp. 153-180.

Other References

Bledsoe, W. W. and Browning, I , "Pattern Recognition and Reading by Machine," Proceedings of the Eastern Joint Computer Conference, December 1959, pp. 225-232.

Chow, C. K., "An Optimum Character Recognition System Using Decision Functions," Institute of Radio Engineers Transactions on Electronic Computers, Vol. EC-6, 1957, p. 247.

Diamond, T. L., "Devices for Reading Handwritten Characters," Proceedings of the Eastern Joint Computer Conference, 1957, pp. 232-237.

Fitzmaurice, John A., Sabbagh, Edward N., and Elliot, William G., "Optical Filter Text Reader Study," ASTIA Document No. AD-217675L, August 20, 1959.

Flores, I., and Grey, L., "Optimization of Reference Signals for Character Recognition Systems," Institute of Radio Engineers Transactions on Electronic Computers, Vol. EC-9, No. 1, March 1960, pp. 54-61.

Greanias, E. C., Hoppel, E. J., Kloomok, M., and Osborne, J. S., "The Design of the Logic for the Recognition of Printed Characters by Simulation," Proceedings Institute of Electrical Engineers, Vol. 103B, Suppl. 3, November 1956, p. 456.

Highleyman, W. H., "Linear Decision Functions, with Application to Pattern Recognition," Proceedings of the Institute of Radio Engineers, Vol. 50, No 6, Pt. 1, June 1962, p. 1501.

Minsky, M. L., "A Selected Descriptor-Indexed Bibliography to the Literature on Artificial Intelligence," Institute of Radio Engineers Transactions on Human Factors, Vol. HFE-2, March 1961, pp. 39-55.

Sprick, W., and Ganshorn, K., "Recognition of Numerals by Contour Following," Proceedings of the Institution of Electronic Engineers, Vol. 106, Pt. A, Section 1, February 16-17, 1959.

Stearns, S. D., "Method for Design of Pattern Recognition Logic," Institute of Radio Engineers Transactions on Electronic Computers, Vol. EC-9, No. 1, March 1960, pp. 48-53.

Unger, S. H., "Pattern Detection and Recognition," Proceedings of the Institute of Radio Engineers, Vol. 47, No. 10, Pt. 3, October 1959, pp. 1737-1752

Yong, D. A., "Automatic Character Recognition," Electronic Engineer, Vol. 32, January 1960, pp. 2-10.